NUREG/CR-6393
BNL-NUREG-52483

Integrated System Validation: Methodology and Review Criteria

Prepared by
J. O'Hara, W. Stubler, J. Higgins, and W. Brown

Brookhaven National Laboratory

Prepared for
U.S. Nuclear Regulatory Commission

NUREG/CR-6393
BNL-NUREG-52483
RX

Integrated System Validation: Methodology and Review Criteria

Manuscript Completed: September 1995
Date Published:

Prepared by
J. O'Hara, W. Stubler, J. Higgins, and W. Brown

Brookhaven National Laboratory, Upton, NY 11973

Prepared for
Human Factors Assessment Branch
Office of Nuclear Reactor Regulation
U.S. Nuclear Regulatory Commission
Washington, DC 20555
NRC FIN E2090

ABSTRACT

This technical report (TR) has been prepared by Brookhaven National Laboratory for the Human Factors Assessment Branch of the U.S. Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation. This report is submitted under the *Advanced Reactor Human Factors Review Project* (FIN E-2090) as part of Task 4, "HFE Program Review Model." The NRC Project Manager is Karen Pulsipher and the Project Engineer is Garmon West. The BNL Principal Investigator is John O'Hara.

CONTENTS

CONTENTS (cont.)

Page

CONTENTS (cont.)

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The U.S. Nuclear Regulatory Commission (NRC) reviews the human factors engineering (HFE) aspects of advanced nuclear power plant designs. In order to support the advanced reactor design certification reviews, the HFE Program Review Model (HFE PRM) (U.S. NRC, 1994) was developed. The HFE PRM describes the HFE program elements that are necessary and sufficient to develop an acceptable detailed design specification and an acceptable implemented design and provides the review criteria for their evaluation. One of the review elements is verification and validation (V&V). The role of V&V evaluations in the HFE PRM is to comprehensively determine that the design conforms to HFE design principles and that it enables plant personnel to successfully perform their tasks to achieve plant safety and other operational goals. Integrated system validation is part of this review activity. However, the HFE PRM provides general criteria for the review of integrated system validation at a program plan level of detail and does not provide sufficient criteria for the review of validation implementation plan methodology and the results of validation tests. The purpose of this document is to more clearly define the detailed methodological considerations necessary for the detailed review of a nuclear power plant (NPP) HFE validation.

The literature associated with the test, evaluation, and validation of complex systems was reviewed. A complex human-machine system may be characterized as one which supports a dynamic process involving many elements that interact in ways that may not be anticipated by the designer. These characteristics are likely to pose significant cognitive demands on operators, both individually and as a crew. Historically, systems have been "validated" when the reliability of their components have been demonstrated. However, this approach to evaluating the acceptability of complex systems is inadequate because their performance is an emergent property out of the integration of all the components, and not simply a product of them. Thus, an evaluation "paradigm" to accomplish integrated system validation is needed.

A paradigm is defined as an example serving as a model or pattern. The paradigm provides a conceptual approach to validation by identifying important validation principles and their relationships. The general concepts in the paradigm are concerned with (1) establishing the requirements for making a logical and defensible inference from validation tests to predicted integrated system performance under actual operating conditions, and (2) identifying the aspects of validation methodology that are important to the inference process. While it is recognized that differences in specific methodologies are possible, the general principles and concepts that are described by the paradigm are invariant across methodologies. The integrated system validation paradigm was developed using (1) the existing HFE PRM review criteria; (2) system test, evaluation, and validation literature; and (3) principles adopted from scientific research methodology.

Once the validation paradigm was identified, considerations were made as to the methodological aspects of the validation process that are needed to meet the general paradigm requirements. That is, while the paradigm identifies the requirements of the inference process, the next task was to identify a means by which the paradigm requirements can be satisfied. Based upon the detailed methodological considerations, criteria were then developed that would enable one to review either an HFE integrated system validation plan or the results of an actual validation program. The criteria also will allow one to identify any weaknesses or threats to the inference process that is necessary for the validation. A high-level overview of the validation process follows.

The objective of validation is to provide evidence that the integrated system adequately supports plant personnel in the safe operation of the plant; i.e., that the integrated design remains within acceptable performance envelopes. To accomplish this objective, the methodology must permit a logical and defensible inference to be made from validation tests to predicted integrated system performance under actual operating conditions. The validation paradigm is based upon four general forms of validity. *System representation validity* refers to the degree to which the validation tests include those aspects of the integrated system that are important to real-world conditions. Specifically, this validity is based on the representativeness of the system model, human-system interface, personnel, and operational events. The inference process is supported to the extent that important aspects of the integrated system are represented with high fidelity, and to the extent to which important contributors to potential system performance variability have been adequately sampled. *Performance representation validity* refers to the completeness and representativeness of the performance measures. A comprehensive, hierarchal approach to evaluation guided by supervisory control theory may be used to specify important aspects of performance ranging from operator cognitive processes to system functions. Failure to include measures of all important performance variables, poor measurement properties, and poor criteria specification weaken this validity. *Test design validity* addresses the procedures used for the conduct of the tests. Inappropriate test procedures can bias the relationship between the observations of performance and the integrated system, and thus undermine their causal linkage. When factors introduced by the test methodology weaken the ability to interpret the system-performance correlation, validity is compromised. *Statistical conclusion validity* addresses the relationship between the observed data and established performance criteria, and, later, the inference from the observed sample to actual performance.

An important aspect of validating an integrated system is establishing that these four types of validity are satisfied. Such assessments are made by reviewing the methodology used to conduct validation tests. Methodological factors relevant to each of the aspects of validity identified above are discussed in the document.

The limitations to integrated system validation are discussed as well. While limitations to integrated system validation are recognized and discussed, it is important to emphasize that a fundamental principle of the HFE PRM is that the complete safety evaluation is based upon the establishment of convergent validation across different evaluation methodologies, each with their strengths and limitations.

ACRONYMS

AIAA	American Institute of Aeronautics and Astronautics
ANS	American Nuclear Society
ANSI	American National Standards Institute
BWR	boiling water reactor
CR	control room
CSF	critical safety function
DoD	U.S. Department of Defense
ECG	electrocardiogram
EEG	electroencephalographic
EOF	emergency offsite facility
EOG	electro-oculogram
EOP	emergency operating procedure
EP	evoked potential
EPRI	Electric Power Research Institute
HFE	human factors engineering
HFE PRM	Human Factors Engineering Program Review Model
HPCI	high pressure core injection
HPM	human performance measurement
HRA	human reliability analysis
HSI	human-system interface
Hz	hertz
I&C	instrumentation and control
IAEA	International Atomic Energy Agency
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
LCO	limiting condition for operation
LCS	local control station
msec	millisecond
MUX	multiplexer
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NPP	nuclear power plant
NRC	U.S. Nuclear Regulatory Commission
PORV	Power-Operated Relief Valve
PRA	probabilistic risk assessment
PSF	performance shaping factor
PWR	pressurized water reactor
PZR	pressurizer
RCIC	reactor core isolation cooling system
RCS	reactor coolant system
RHR	residual heat removal
RPM	revolutions per minute
RPV	reactor pressure vessel
RVLIS	reactor vessel level indication systems
Rx	reactor

ACRONYMS (Cont'd.)

SA	situation awareness
SA-SWORD	Situation Awareness Subjective Workload Dominance
SAGAT	Situation Awareness Global Assessment Technique
SAR	safety analysis report
SART	Situation Awareness Rating Technique
SME	subject matter expert
SP	suppression pool
SPDS	Safety Parameter Display System
SWAT	Subjective Workload Assessment Technique
SWORD	Subjective Workload Dominance
TLX	Task Load Index
TMI	Three Mile Island
TSC	technical support center
V&V	verification and validation

GLOSSARY

Bias - Bias is an aspect of the methodology which systematically modifies performance.

Cognitive error - A human error that results from the characteristics of human information processing such as errors in diagnosis due to information overload.

Complex Human-Machine System - A complex human-machine system may be defined as one supporting a dynamic process involving a large number of elements that interact in many different ways.

Component - An individual piece of equipment such as a pump, valve, or vessel; usually part of a plant system.

Confound - A confound is the systematic coupling of one aspect of the test with another aspect of the test or an extraneous variable. Confounding makes important relationships ambiguous.

Construct Validity - The extent to which a selected performance measure accurately represents of the aspect of performance one wants to measure.

Convergent Validity - Convergent validity is the degree to which consistent results are observed across different review, evaluation, or measurement techniques.

Critical tasks - Tasks that must be accomplished in order for personnel to perform their functions. In the context of probabilistic risk assessment, critical tasks are those that are determined to be significant contributors to plant risk.

Function - An action that is required to achieve a desired goal. Safety functions are those functions that serve to ensure higher-level objectives and are often defined in terms of a boundary or entity that is important to plant integrity and the prevention of the release of radioactive materials. A typical safety function is "reactivity control." A high-level objective, such as preventing the release of radioactive material to the environment, is one that designers strive to achieve through the design of the plant and that plant operators strive to achieve through proper operation of the plant. The function is often described without reference to specific plant systems and components or the level of human and machine intervention that is required to carry out this action. Functions are often accomplished through some combination of lower-level functions, such as "reactor trip." The process of manipulating lower-level functions to satisfy a higher-level function is defined here as a control function. During function allocation the control function is assigned to human and machine elements.

Human-centered design goals - Human factors engineering design goals that address the cognitive and physical support of personnel performance.

Human factors - A body of scientific facts about human characteristics. The term covers all biomedical, psychological, and psychosocial considerations; it includes, but is not limited to, principles and applications in the areas of human factors engineering, personnel selection, training, job performance aids, and human performance evaluation (see "Human factors engineering").

Human factors engineering (HFE) - The application of knowledge about human capabilities and limitations to plant, system, and equipment design. HFE ensures that the plant, system, or equipment design, human tasks, and work environment are compatible with the sensory, perceptual, cognitive, and physical attributes of the personnel who operate, maintain, and support it (see "Human factors").

Human-system interface (HSI) - The means through which personnel interact with the plant, including the alarms, displays, controls, and job performance aids. Generically this includes maintenance, test, and inspection interfaces as well.

Integrated System Validation - The HFE PRM states that the purpose of integrated system validation is to provide evidence that the integrated system adequately supports plant personnel in the safe operation of the plant; i.e., that the integrated design remains within acceptable performance envelopes.

Local control station (LCS) - An operator interface related to nuclear power plant (NPP) process control that is not located in the main control room. This includes multifunction panels, as well as single-function LCSs such as controls (e.g., valves, switches, and breakers) and displays (e.g., meters) that are operated or consulted during normal, abnormal, or emergency operations.

Masking - Masking is the addition of noise or error variance to performance data, which makes the results more difficult to interpret and the prediction of actual plant performance less certain.

Mockup - A static representation of an HSI (see "Simulator" and "Prototype").

Paradigm - An example that serves as a model or pattern.

Performance Representation Validity - Performance representation validity refers to the degree to which performance measures adequately represent those performance characteristics that are important to safety. Thus performance representation validity is supported when a measure is representative of the aspect of performance to be measured.

Performance shaping factors (PSFs) - Factors that influence human reliability through their effects on performance. PSFs include factors such as environmental conditions, HSI design, procedures, training, and supervision.

Personal Safety - See "Safety."

Plant - The nuclear power plant in its entirety including all plant systems and components.

Plant Safety - See "Safety."

Primary Tasks - Primary tasks are those involved in performing the functional role of the operator to supervise the plant; i.e., process monitoring, decision-making, and control.

Prototype - A dynamic representation of an HSI that is not linked to a process model or simulator (see "Simulator" and "Mockup").

GLOSSARY (Cont'd.)

Safety - The term used in the following contexts in the HFE Program Review Model:

Personal safety - Relates to the prevention of individual accidents and injuries of the type regulated by the Occupational Safety and Health Administration.

Plant safety - Also called "safe operation of the plant." A general term used herein to denote the technical safety objective as articulated by the International Nuclear Safety Advisory Group of the International Atomic Energy Agency (IAEA) in the "Basic Safety Principles for . . Nuclear Power Plants" (IAEA, 1988): "To prevent with high confidence accidents in nuclear . . plants; to ensure that, for all accidents taken into account in the design of the plant, even . . . . . those of very low probability, radiological consequences, if any, would be minor; and to . . . . . ensure that the likelihood of severe accidents with serious radiological consequences is . . . . . extremely small." See Section 1.4 for additional discussion.

Safety evaluation - The NRC process of reviewing an aspect of an NPP to ensure that it meets requirements and that it will perform as needed to reliably ensure plant safety.

Safety function - See "Function."

Safety issue - An item identified during plant design, operation, or review that has the . potential to affect the safe operation of the plant.

Safety-related - A term applied to those NPP structures, systems, and components (SSCs) that prevent or mitigate the consequences of postulated accidents that could cause undue risk . . . to the health and safety of the public (see Appendix B to Part 50 of Title 10 of the U.S. Code . . . . . of Federal Regulations). These are the SSCs on which the design-basis analyses of the safety . . . . analysis report are performed. They also must be part of a full quality assurance program in . . accordance with Appendix B.

Secondary Tasks - Secondary tasks are those the operator must perform when interfacing with the plant, but which are not directed to the primary task, e.g., navigating through and paging displays, searching for data, choosing between multiple ways of accomplishing the same task, and making decisions regarding how to configure the interface.

Simulator - A facility that physically represents the HSI configuration and that dynamically represents the operating characteristics and responses of the plant in real time (see "Prototype" and "Mockup").

Situation awareness - The relationship between the operator's *understanding* of the plant's condition and its actual condition at any given time.

Statistical Conclusion Validity - Statistical conclusion validity addresses the relationship between the performance data and the established performance criteria.

GLOSSARY (Cont'd.)

Subsidiary Tasks - Subsidiary tasks are those used for workload assessment. These tasks are given to operators to perform while they are performing primary and secondary tasks. Their performance is theoretically tied to spare mental capacity approaches to cognitive workload. Better performance on subsidiary tasks reflects more spare mental capacity and, therefore, lower primary/secondary task workload.

System - An integrated collection of plant components and control elements that operate alone or with other plant systems to perform a function.

System Representation Validity - System representation validity refers to the degree to which the integrated system validation tests include those aspects of the integrated system that are important to real-world conditions, including the process/plant model, human-system interface (HSI), plant personnel, and plant operational conditions.

Task - A group of activities that have a common purpose, often occurring in temporal proximity, and that utilize the same displays and controls.

Test Design Validity - Test design validity addresses those considerations that are involved in the actual conduct of the validation tests, including activities such as the assignment of crews to scenarios, development of test procedures, and participant training.

Top-down design - A review approach starting at the "top" with high-level plant mission goals that are decomposed into functions that are allocated to human and system resources and are decomposed into tasks required to accomplish function assignments. Tasks are arranged into meaningful jobs and the HSI is designed to best support job task performance. The detailed design is the "bottom" of the top-down process.

Type I error - Reflects an incorrect decision that the design is acceptable, a false positive.

Type II error - Reflects an incorrect decision that the design is unacceptable, a false negative.

Validation - Describes a process by which integrated system design (consisting of hardware, software, and personnel elements) is evaluated to determine whether it acceptably supports safe operation of the plant.

Validity - See the specific uses of the term: construct, convergent validity, performance representation validity, statistical conclusion validity, system representation validity, and test design validity.

Vigilance - The degree to which personnel are attentive to their current task.

Workload - The physical and cognitive demands placed on plant personnel.

Part I

Integrated System Validation

Review Criteria

1. INTRODUCTION

1.1 NRC Human Factors Review of Advanced Reactor Designs

The U.S. Nuclear Regulatory Commission (NRC) reviews the human factors engineering (HFE) aspects of advanced nuclear power plant (NPP) designs to ensure that they are designed to good HFE principles and that operator performance and reliability are appropriately supported in order to protect public health and safety. In order to support the advanced reactor design certification reviews, the NRC, in conjunction with Brookhaven National Laboratory, has developed an HFE Program Review Model (HFE PRM, O'Hara et al., 1994). The HFE PRM describes the HFE program elements that are necessary and sufficient to develop an acceptable detailed design specification and an acceptable implemented design and provides the review criteria for their evaluation. The HFE PRM is being improved through the development of additional review procedures in selected areas. One such area is integrated system validation which is the subject of this document. The role of HFE validation in the evaluation of plant safety is briefly discussed below.

Plant safety, also called "safe operation of the plant," is a general term used herein to denote the technical safety objective as articulated by the International Atomic Energy Agency (IAEA):

> "To prevent with high confidence accidents in nuclear plants; to ensure that, for all accidents taken into account in the design of the plant, even those of very low probability, radiological consequences, if any, would be minor; and to ensure that the likelihood of severe accidents with serious radiological consequences is extremely small" (IAEA, 1988).

To ensure plant safety requires "defense in depth." Defense in depth includes the use of multiple barriers to prevent the release of radioactive materials and uses a variety of programs to ensure the integrity of barriers and related systems [a detailed discussion of this approach is provided in the IAEA basic safety principles (IAEA, 1988)]. These programs include, among others, conservative design, quality assurance, administrative controls, safety reviews, personnel qualification and training, test and maintenance, safety culture, and human factors.

The NRC process of reviewing an aspect of an NPP to ensure that it meets requirements and that it will perform as needed to reliably ensure plant safety is called a "safety evaluation." The HFE PRM provides a top-down approach for the conduct of an NRC safety evaluation of an NPP HFE program. Top-down refers to a review approach starting at the "top" with high-level plant mission goals that are broken down into the functions necessary to achieve the mission goals. Functions are allocated to human and system resources and are broken down into tasks for the purposes of specifying the alarms, information, and controls that will be required to accomplish function assignments. Tasks are arranged into meaningful jobs and the HSI is designed to best support job task performance. The detailed design (of the HSI, procedures, and training) is the "bottom" of the top-down process. The HFE safety evaluation is broad based and includes HFE aspects of normal and emergency operations, test, maintenance, etc.

The rationale underlying the HFE PRM is that "plant safety" is a concept that is not directly observed but must be inferred from available evidence. When reviewing a design to make a safety evaluation, evidence is collected and weighted toward or against an acceptable finding. As in the assessment of any inferred concept, different types of information can be collected. The reviewer seeks to obtain evaluation data from different methods in order to establish "convergent validity" (Campbell and Fisk, 1959), that is, to establish a consistent finding across different types of information, each with its own unique sources of bias and error. This approach to design review is analogous to the defense-in-depth philosophy.

The types of information that can provide assessments of NPP HFE adequacy include:

- HFE planning (including an HFE design team, program plans, and procedures);
- Design analyses and studies (including requirements, function and task analyses, technology assessments, and tradeoff studies);
- Design specifications and descriptions; and
- Verification and validation (V&V) analyses of the final design (e.g., compliance with accepted HFE guidelines and operation of the integrated system with operators performing the required tasks under actual (or simulated) conditions).

The greatest confidence that a design is acceptable (and ensures plant safety) can be placed in one that has all of the following characteristics: (1) developed by a qualified HFE design team with all the skills required, using an acceptable HFE program plan; (2) resulted from appropriate HFE studies and analyses that provide accurate and complete inputs to the design process and inputs to V&V assessment criteria; (3) designed using proven technology based on human performance and task requirements incorporating accepted HFE standards and guidelines; and (4) evaluated with a thorough V&V test program. Further confidence in the design is then obtained through a detailed initial test program of the actual plant and finally through successful operation over a period of time.

Similar approaches to complex system evaluation are emerging in other industries (e.g., Miller et al., 1994). With regard to design certification for civil aviation systems, Stager (1994) has stated that "the primary objectives of human factors certification must be accomplished within the design and validation phases of the human engineering program and that human factors certification of more complex cognitive systems is tantamount to certification of the underlying design development methodology" (p. 1055).

1.2 Integrated System Validation in the HFE PRM

The role of V&V evaluations in the HFE PRM is to comprehensively determine that the design conforms to HFE design principles and that it enables plant personnel to successfully perform their tasks to achieve plant safety and other operational goals. The HFE PRM V&V element is made up of the five activities, including:

- HSI Task Support Verification - a check to ensure that HSI components are provided to address all identified personnel tasks

- HFE Design Verification - a check to determine whether the design of each HSI component reflects HFE principles, standards, and guidelines

- Integrated System Validation - performance-based evaluations of the integrated design to ensure that the HFE/HSI supports safe operation of the plant.

- Human Factors Issue Resolution Verification - a check to ensure that the HFE issues identified during the design process have been acceptably addressed and resolved.

- Plant HFE Verification - the "final" design should be documented in a design description document that includes the requirements for verification that the "as built" design is the same as the design resulting from the design process V&V evaluations. This document can then be used to conduct a final plant HFE/HSI design verification. The main activity should be a check of the actual HSIs against the description.

As indicated above, the purpose of integrated system validation is to provide evidence that the integrated HSI adequately supports operating crew performance in the safe operation of the plant; i.e., that the integrated design can perform within an acceptable performance envelope. The HFE PRM identifies general criteria for the evaluation of validation in Section 11.4.4, including (some of the review criteria have been abbreviated for the discussion below):

1. The methodology for integrated system validation should address: general objectives, personnel performance issues to be addressed, test methodology and procedures, test participants, test conditions, HSI description, performance measures, data analysis, criteria for evaluation of results, and utilization of evaluations.

2. Validation should be performed by evaluating dynamic task performance using tools that are appropriate to the accomplishment of this objective. The primary tool for this purpose is a simulator, that is, a facility that physically represents the HSI configuration and that dynamically represents the operating characteristics and responses of the plant design in real time. The requirement to validate performance at plant HSIs outside the control room (CR) will be dependent on the applicant's design. Human actions at non-CR facilities such as remote shutdown panels and local control stations may be evaluated using mockups, prototypes, or similar tools.

3. The evaluations should address

- adequacy of the entire HSI configuration for achievement of HFE program goals
- confirmation of allocation of function and the structure of tasks assigned to personnel
- adequacy of staffing and the HSI to support staff to accomplish their tasks
- adequacy of procedures
- confirmation of the dynamic aspects of the HSI for task accomplishment
- evaluation and demonstration of error tolerance to human and system failures

4. All critical human actions as defined by the task analysis and probabilistic risk analysis/human reliability analysis (PRA/HRA) should be tested and found to be adequately supported in the design, including the performance of critical actions outside the control room. The design of tests and evaluations to be performed as part of HFE V&V activities should specifically examine these actions.

5. The validation should evaluate selected activities based on procedures developed to address Regulatory Guide 1.33, Appendix A category procedures.

6. Dynamic evaluations should evaluate the HSI under a range of operational conditions and upsets, and should include the following: normal plant evolutions, instrument failures, HSI equipment and processing failure, transients, accidents, and reactor shutdown and cooldown from the remote shutdown panel.

7. The scenarios should be realistic. Selected ones should include environmental conditions such as noise and distractions that may affect human performance in an actual nuclear power plant. For actions outside the CR, the performance impacts of potentially harsh environments (i.e., high radiation) that require additional time should be realistically simulated (i.e., time to don protective clothing and access hot areas).

8. Performance measures for dynamic evaluations should be adequate to test the achievement of all objectives, design goals, and performance requirements and should include the following at a minimum: system performance measures relevant to plant safety, primary task performance and errors, situation awareness, workload, crew communications and coordination, dynamic anthropometry evaluations, and physical positioning and interactions.

1.3 Objectives, Use, and Document Organization

The HFE PRM provides general criteria for the review of integrated system validation at a program plan level of detail. However, at present it does not provide sufficient criteria for the review of validation implementation plan methodology and the results of validation tests. The purpose of this document is to more clearly define the detailed methodological considerations necessary for the detailed review of an NPP HFE validation.

The document is divided into two parts. Part I presents the integrated system validation review criteria. The detailed criteria are contained in Section 2.

Part II, Criteria Development and Technical Basis, documents the approach to validation upon which the criteria are based. Section 3 describes general methodology and bases upon which the review criteria were developed. Section 4 describes the development of a general validation paradigm; i.e., a conceptual approach to validation, its important validity principles, and their relationships. Section 5 describes the considerations for meeting the requirements of the paradigm.

Table 1.1 provides links between the eight validation review criteria identified in Section 1.2 and the information in this document. While conducting an integrated system validation review, the reviewer can consult the information in Sections 4 and 5 to support the evaluation of the identified HFE PRM criterion.

It is important to note that this document addresses additional topics not specifically addressed in the HFE PRM. These topics are identified in Table 1.2.

Table 1.1 HFE PRM Validation Review Criteria and Sections of This Document

HFE PRM Criteria	Section 2 Review Criteria	Section 4 Paradigm	Section 5 Methodology
1 - Method Topics	2 (all)	--	5 (all)
2 - Testbeds	2.3	4.2.2	5.3
3 - Objectives	2.2	4.1	5.2
4 - Critical Actions	2.5	4.2.2	5.5
5 - Procedures	2.5	4.2.2	5.5
6 - Scenarios	2.5	4.2.2	5.5
7 - Realism	2.5	4.2.2	5.5
8 - Performance Measures	2.6	4.2.3	5.6

Table 1.2 New Validation Review Topics

New Criteria Topics	Section 2 Review Criteria	Section 4 Paradigm	Section 5 Methodology
Validation Team	2.1	--	5.1
Test Participants	2.4	4.2.2	5.4
Test Design	2.7	4.2.3	5.7
Data Analysis/Interpretation	2.8	4.2.4	5.8
Conclusions	2.9	4.2.1, 4.3, 4.4	5.9

2. REVIEW CRITERIA

This section provides criteria addressing validation methodology, including:

- Validation Team (Section 2.1)
- Test Objectives (Section 2.2)
- Validation Testbeds (Section 2.3)
- Plant Personnel (Section 2.4)
- Operational Conditions (Section 2.5)
- Performance Measurement (Section 2.6)
- Test Design (Section 2.7)
- Data Analysis and Interpretation (Section 2.8)
- Validation Conclusion (Section 2.9)

The review criteria are presented in the format used in the HFE PRM. Table 1.1 provides the relationship between the HFE PRM review criteria for integrated system validation and the criteria presented in this section.

The criteria are based on concepts and technical bases that are discussed in Part II of this report. The explanations and discussions of the concepts are not included in the criteria below. Therefore, familiarity with that material is necessary to gain a full understanding of the criteria below.

2.1 Validation Team

(1) The validation team should be multidisciplinary. Appropriate areas of expertise are described in Appendix A of the HFE PRM. Each of the technical disciplines listed in the HFE PRM may not be necessary. Rather, the specific technical areas of expertise required for the validation team should be based on the scope of the validation effort. In addition to the skills listed in the HFE PRM appendix, the validation team should include personnel with expertise in test and evaluation, including test design, test procedure development, performance measurement, and data analysis.

(2) The members of the validation team should have independence from the personnel responsible for the actual design.

2.2 Test Objectives

(1) Detailed objectives should be developed to provide evidence that the integrated system adequately supports plant personnel in the safe operation of the plan. The objectives should include:

- Validate the role of plant personnel

- Validate that the shift staffing, assignment of tasks to crew members, and crew coordination (both within the control room as well as between the control room and local control stations and support centers) is acceptable. This should include validation of the nominal shift levels, minimal shift levels, and shift turnover.

- Validate that for each human function, the design provides adequate alerting, information, control, and feedback capability for human functions to be performed under normal plant evolutions, transients, design basis accidents, and selected, risk-significant events that are beyond-design basis.

- Validate that specific personnel tasks can be accomplished within time and performance criteria, with a high degree of operating crew situation awareness, and with acceptable workload levels that provide a balance between vigilance and operator burden. Validate that the operator interfaces minimize operator error and provide for error detection and recovery capability when errors occur.

- Validate that the functional requirements are met for the major HSI features, e.g., group-view display, alarm system, safety parameter display system (SPDS) function, general display system, procedures, controls, communication systems, controls EOP-related local control stations.

- Validate that the crew can make effective transitions between the HSI features in the accomplishment of their tasks and that interface management tasks such as display configuration and navigation are not a distraction or undue burden.

- Validate that the integrated system performance is tolerant of failures of individual HSI features.

- Identify aspects of the integrated system (including staffing, communications, and training) that may negatively impact integrated system performance.

(2) Detailed objectives should be defined in a systematic manner which relates scenario characteristics and performance measurement criteria.

2.3 Validation Testbeds

The criteria for testbeds are divided into three sections. Section 2.3.1 addresses characteristics of the main control room, Section 2.3.2 addresses the representation monitoring and control facilities remote from the main control room, and Section 2.3.3 addresses testbed verification prior to conducting validation trials.

2.3.1 Main Control Room

(1) HSI completeness - The HSI should be completely represented in the testbed. This should also include HSI not specifically required in the test scenarios.

(2) HSI Physical Fidelity - A high degree of physical fidelity in the HSI should be represented, including presentation of alarms, displays, controls, job aids, procedures, communications, interface management tools, layout and spatial relationships.

(3) HSI Functional Fidelity - A high degree of functional fidelity in the HSI should be represented. All HSI functions should be available. High functional fidelity includes HSI component modes of operation, i.e., the changes in functionality that can be invoked on the basis of operator selection and/or plant states.

(4) Data Completeness Fidelity - Information and data provided in the control room should completely represent the plant systems monitored and controlled from that facility.

(5) Data Content Fidelity - A high degree of data content fidelity should be represented. The information and controls presented at the HSI should be based on an underlying model that accurately reflects the reference plant. The model should provide input to the HSI in a manner such that information accurately matches that which will be presented in the actual control room.

(6) Data Dynamics Fidelity - A high degree of data dynamics fidelity should be represented. The process model should be capable of providing input to the HSI in a manner such that information flow and control responses occur accurately and in a response time that matches that in the actual control room. Overall, the HSI should provide the same response times as the actual control room; e.g., information should be provided to the operator with the same delays as would occur in the plant.

(7) Environment Fidelity - A high degree of environment fidelity should be represented. The lighting and noise characteristics of the control room should reasonably reflect that expected in the actual control room. Thus, noise contributed by equipment, such as air handling units, and computers, etc. should be represented in validation tests.

2.3.2 Representation Monitoring and Control Facilities Remote from the Main Control Room

(1) For important actions at complex HSIs remote from the main control room, where timely and precise human actions are required, the use of a simulation or mockup to verify that human performance requirements can be achieved should be considered. (For less critical actions or where the HSI are not complex, human performance may be assessed based on analysis rather than simulation.)

(2) When simulations or mockups are used, the important characteristics of the task-related HSIs and task environment (e.g., lighting, noise, heating and ventilation, and protective clothing and equipment) should be included in the testbed.

2.3.3 Testbed Verification

(1) The testbed should be verified for conformance to the testbed characteristics identified in 2.3.1 above prior to validation.

2.4 Plant Personnel

(1) Participants in validation test should be representative of actual plant personnel who will interact with the HSI, e.g., active operators rather than training or engineering personnel.

(2) To properly account for human variability, a sample of participants should be used. The sample should reflect the characteristics of the population from which the sample is drawn. Those characteristics that are expected to contribute to system performance variation should be specifically identified and the sampling process should ensure that variation along that dimension is included in the validation. Several factors that should be considered in determining representativeness include: license and qualifications, skill/experience, age, and general demographics.

(3) Shift Staffing - In selection of personnel, consideration should be given to the assembly of operating crews, e.g., shift supervisors, reactor operators, shift technical advisors, etc., that will participate in the tests.

(4) To prevent bias in the sample, the following participant characteristics and selection practices should be avoided:

- Participants who are part of the design organization
- Participants in prior evaluations
- Participants who are selected for some specific characteristic, such as using crews that are identified as good or experienced.

2.5 Operational Conditions

The criteria for operational conditions are divided into two sections. Section 2.5.1 addresses the operational conditions sampling and Section 2.5.2 addresses scenario definition.

2.5.1 Operational Conditions Sampling

(1) Integrated system validation should include dynamic evaluations for a range of operational conditions that are representative of actual plant conditions. A sample of operational conditions should be used that are important to safety, and should include conditions that are representative of the range of events that could be encountered during operation of the plant. The sample should reflect the characteristics of the population from which the sample is drawn. Those characteristics that are expected to contribute to system performance variation should be specifically identified and the sampling process should ensure that variation along that dimension is included in the validation. The sampling dimensions are addressed in criteria 2, 3, and 4).

(2) *Plant Conditions* - The validation scenarios should include the following:

- Normal operational events including plant startup, plant shutdown or refueling, and significant changes in operating power.

- Failure events such as:

  - Instrument failures (e.g., safety-related system logic and control unit, fault tolerant controller, local "field unit" for multiplexer (MUX) system, MUX controller, and break in MUX line) including I&C failures that exceed the design basis, such as a common mode I&C failure during an accident.

  - HSI failures (e.g., loss of processing and/or display capabilities for alarms, displays, controls, and computer-based procedures).

- Transients and accidents such as:

  - Transients (e.g., turbine trip, loss of off-site power, station blackout, loss of all feedwater, loss of service water, loss of power to selected buses or CR power supplies, and safety and relief valve transients).

  - Accidents (e.g., main steam line break, positive reactivity addition, control rod insertion at power, anticipated transient without scram, and various-sized loss-of-coolant accidents).

  - Reactor shutdown and cooldown using the remote shutdown system.

- Reasonable, risk-significant, beyond-design-basis events.

  - These should be determined from the plant specific probabilistic risk assessment (PRA).

- In selecting failures, consideration should be given to the role of the equipment in achieving plant safety functions (as described in the plant SAR) and the degree of interconnection with other plant systems. A system that is interconnected with other systems could cause the failure of other systems because the initial failure could propagate over the connections. This consideration is especially important when assessing non-class 1E electrical systems.

(3) *Personnel Tasks* - The scenario should reflect a range of interactions with HSI components and personnel:

- Range of risk-significant actions, systems, and accident sequences - The scenarios should test all risk-important human actions as defined by the task analyses and PRA and HRA, including those performed outside the control room. Situations where human monitoring of an automatic system is critical should be considered. Additional factors that contribute highly to risk, as defined by the PRA, should be sampled, including:

  - Dominant human actions (selected via sensitivity analyses),

  - Dominant accident sequences, and

  - Dominant systems (selected via PRA importance measures such as Risk Achievement Worth or Risk Reduction Worth).

- Range of procedure guided tasks - Regulatory Guide 1.33, Appendix A, contains several categories of "typical safety-related activities that should be covered by written procedures." The validation should evaluate selected activities based on procedures developed to address this guide. The evaluation should include appropriate procedures in each relevant category:

  - Administrative procedures
  - General plant operating procedures
  - Procedures for startup, operation, and shutdown of safety-related systems
  - Procedures for abnormal, offnormal, and alarm conditions
  - Procedures for combating emergencies and other significant events
  - Procedures for control of radioactivity
  - Procedures for control of measuring and test equipment and for surveillance tests, procedures, and calibration
  - Procedures for performing maintenance
  - Chemistry and radiochemical control procedures

- Range of human decision-making activities - The scenarios should reflect the range of activities performed by personnel, including:

  - Monitoring and detection (e.g., of critical safety-function threats),
  - Interpretation/diagnosis (e.g., interpretation of alarms and displays for diagnosis of faults in plant processes and automated control and safety systems),
  - Planning (e.g., evaluating alternatives for recovery from plant failures),
  - Execution (e.g., In-the-loop control of plant systems, assuming manual control from automatic control systems, and carrying out complicated control actions),
  - Obtaining feedback (e.g., of the success of actions taken).

  The range of scenarios should include tasks that exemplify skill, rule, and knowledge-based behavior.

- Range of HSI components - The scenarios should address use of all types of HSI components:

  - Alarm system,
  - Display systems (e.g., discrete indicators, process displays, group-view displays),

- Control systems: manual, automated, and combined manual and automated,
- Interface management facilities such as dialog design and navigation,
- Procedures,
- Job support and decision aids, and
- Communication equipment.

- Range of human interactions - The scenarios should reflect the range of interactions between plant personnel, including tasks that are performed independently by individual crew members and tasks that are performed by crew members acting as a team. These interactions between plant personnel should include:
  - Between main control room operators (e.g., operations, shift turnover walkdowns),
  - Main control room operators and auxiliary operators,
  - Main control room operators and support centers (e.g., the technical support center and the emergency offsite facility), and
  - Main control room operators with plant management, NRC, and other outside organizations.
- Tasks that are performed with high frequency.

(4) *Situational factors that are known to challenge human performance* - The scenario should reflect a range of situational factors that are known to challenge human performance, such as:

- Difficult NPP Tasks - The scenarios should address tasks that have been found to be problematic in the operation of NPPs, e.g., procedure versus situation assessment conflicts. The specific tasks selected should reflect the operating history of the type of plant being validated (or the plant's predecessor).
- Error-forcing contexts - Situations specifically design to create human errors should be included in validation to assess the error tolerance of the system and the capability of operators to recover from errors should they occur.
- The scenarios should include situations where human performance variation due to high workload and multitasking situations can be assessed.
- The scenarios should include situations where human performance variation due to workload transitions can be assessed. These include conditions that exhibit (1) a sudden increase in the number of signals that must be detected and processed following a period in which signals were infrequent and (2) a rapid reduction in signal detection and processing demands following a period of sustained high task demand.

- The scenarios should include situations where human performance variation due to personnel fatigue and circadian factors can be assessed.

- The scenarios should include situations where human performance variation due to environmental conditions such as poor lighting, extreme temperatures, and high noise can be assessed.

(5) The sample should not be biased in the direction of over representation of the following:

- Scenarios for which only positive outcomes can be expected.

- Scenarios that are relatively easy to conduct (e.g., scenarios that place high demands for simulation, data collection, or analysis are sometimes avoided).

- Scenarios that are familiar and well structured (e.g., which address familiar systems and failure modes that are highly compatible with plant procedures such as "textbook" design-basis accidents).

2.5.2 Scenario Definition

(1) The operational conditions selected for inclusion in the validation tests should be developed into detailed scenarios. The following information should be defined to ensure that important performance dimensions are addressed and to allow scenarios to be accurately presented for repeated trials:

- Description of the scenario mission and any pertinent "prior history" necessary for operators to understand the state of the plant upon scenario start-up

- Specific start conditions (precise definition provided for plant functions, processes, systems, component conditions and performance parameters, e.g., similar to plant shift turnover)

- Events (e.g., failures) to occur and their initiating conditions, e.g., time, parameter values, or events

- Precise definition of workplace factors, such as environmental conditions

- Task support requirements (e.g., procedures and technical specifications)

- Staffing requirements

- Communication requirements with remote personnel (e.g., load dispatcher via telephone)

- Crew behavior requirements (e.g., information gathering, decision making, and plant control actions)

- Data to be collected and the precise specification of what, when and how data is to be obtained and stored (including videotaping requirements, questionnaire and rating scale administrations)

- Specific criteria for terminating the scenario.

(2) Scenarios should have appropriate task fidelity so that realistic task performance will be observed in the tests and so that test results can be generalized to actual operation of the real plant.

(3) When evaluating performance associated with the use of HSI components located remote from the main control room, the effects on crew performance due to potentially harsh environments (i.e., high radiation) should be realistically simulated (i.e., additional time to don protective clothing and access radiologically controlled areas).

2.6 Performance Measurement

The review criteria for performance measurement are divided into three sections. Section 2.6.1 addresses the measurement characteristics that impact the quality of the performance measures, Section 2.6.2 addresses the identification and selection of variables to represent measures of performance, and Section 2.6.3 addresses the development of performance criteria.

2.6.1 Measurement Characteristics

(1) Performance measures should acceptably exhibit the following measurement characteristics (it should be noted that some of the characteristics identified below may not apply to every performance measure):

- Construct validity
- Reliability
- Resolution
- Sensitivity
- Diagnosticity
- Simplicity
- Objectivity
- Impartiality
- Unintrusiveness
- Acceptability
- Administration

2.6.2 Variable Selection

(1) A comprehensive, hierarchal set of performance measures should be used which includes measures of the performance of the plant and personnel (i.e., personnel tasks, situation awareness, cognitive workload, and anthropometric/physiological factors.).

(2) Plant Performance Measurement - plant performance measures representing functions, systems, components, and HSI should be obtained.

(3) Personnel Task Measurement - Two types of personnel tasks should be measured: primary tasks and secondary tasks. Primary tasks are those involved in performing the functional role of the operator to supervise the plant; i.e., process monitoring, decision-making, and control. Secondary tasks are those the operator must perform when interfacing with the plant, but which are not directed to the primary task.

- For each specific scenario, the tasks that personnel *are required to* perform should be identified and assessed. Such tasks can include necessary primary (e.g., start a pump) as well as secondary (e.g., access the pump status display) tasks. This analysis should be used for the identification of errors of omission by identifying tasks which should be performed.

- The tasks that are *actually* performed by personnel during simulated scenarios should be identified and quantified. (Note that the actual tasks may be somewhat different from those that should be performed). Analysis of tasks performed should be used for the identification of errors of commission.

- Primary tasks should be assessed at a level of detail appropriate to the tasks demands. For example, for some simple scenarios, measuring the time to complete a task may be sufficient. For more complicated tasks, especially those that may be described as knowledge-based, it may be appropriate to perform a more fine-grained analysis such as identifying task components: seeking specific data, making decisions, taking actions, and obtaining feedback. Tasks that are critical to successful integrated system performance and are knowledge-based should be measured in a more fine-grained approach.

- The measurement of secondary tasks should reflect the demands of detailed implementation, e.g., time to configure a workstation, navigate between displays, and manipulate displays (e.g., changing display type and setting scale).

- The variable used to quantify tasks should be chosen to reflect the important aspects of the task with respect to system performance, such as:

  - Time
  - Accuracy
  - Frequency
  - Errors (omission and commission)
  - Amount achieved or accomplished
  - Consumption or quantity used
  - Subjective reports of participants

- Behavior categorization by observers

(3) Situation Awareness - Crew and operator situation awareness should be assessed. The approach to situation awareness measurement should be justified.

(4) Cognitive Workload - Crew and operator workload should be assessed. The approach to workload measurement should be justified.

(5) Anthropometric and Physiological Factors - Anthropometric and physiological factors such concerns as visibility of indications, accessibility of control devices, and ease of control device manipulation should be measured where appropriate. Attention should be focussed on those aspects of the design that can only be addressed during testing of the integrated system, e.g., the ability of the operators to effectively use the various controls, displays, workstations, or consoles in an integrated manner.

2.6.3 Performance Criteria

(1) Criteria for the performance measures used in the evaluations should be established.

(2) The approach to establishing criteria should be based upon the type of comparisons between measures and criteria that are performed, e.g., requirement-referenced, benchmark referenced, normative referenced, and expert-judgement referenced.

2.7 Test Design

The review criteria for test design are divided into five sections. Section 2.7.1 addresses coupling crews and scenarios, Section 2.7.2 addresses test procedures, Section 2.7.3 addresses the training of test conductors, Section 2.7.4 addresses the training of test participants, and Section 2.7.5 addresses the conduct of pilot studies.

2.7.1 Coupling Crews and Scenarios

(1) Scenario Assignment - Important characteristics of scenarios should be balanced across crews. Random assignment of scenarios to crews is not recommended. The value of using random assignment to control bias is only effective when the number of crews is quite large. Instead, the validation team should attempt to provide each crew with a similar and representative range of scenarios.

(2) Scenario Sequencing - The order of presentation of scenario types to crews should be carefully balanced to ensure that the same types of scenario are not always being presented in the same linear position, e.g., the easy scenarios are not always presented first.

2.7.2 Test Procedures

(1) Detailed, clear, and objective procedures should be available to govern the conduct of the tests. These procedures should include:

- Information pertaining to the experimental design, i.e., an identification of which crews receive which scenarios and the order that the scenarios should be presented.

- Detailed and standardized instructions for briefing the participants. The type of instructions given to participants can affect their performance on a task. This source of bias can be minimized by developing standard instructions.

- Specific criteria for the conduct of specific scenarios, such as when to start and stop scenarios, when events such as faults are introduced, and the other information discussed in Section 2.5.2, Scenario Definition.

- Scripted responses for test personnel who will be acting as plant personnel during test scenarios. To the greatest extent possible, responses to communications from operator participants to test personnel (serving as surrogate outside the control room personnel) should be prepared. There are limits to the ability to preplan communications since operators may ask questions or make requests that were not anticipated. However, efforts should be made to detail what information personnel outside the control room can provide, and script the responses to likely questions.

- Guidance on when and how to interact with participants when simulator or testing difficulties occur. Even when a high-fidelity simulator is used, the participants may encounter artifacts of the test environment that detract from the performance for tasks that are the focus of the evaluation. Guidance should be available to the test conductors to help resolve such conditions.

- Instructions regarding when and how to collect and store data. These instructions should identify which data are to be recorded by:

  - simulation computers,
  - special purpose data collection devices (such as automated situation awareness data collection, workload measurement, or physiological measures),
  - video recorders (locations and views),
  - test personnel in real time (such as observation checklists), and
  - subjective rating scales and questionnaires.

- Instructions for maintaining and updating test conductor logs. These instructions should detail the types of information that should be logged (e.g., when tests were performed, deviations from test procedures, and any unusual events that may be of importance to understanding how a test was run or interpreting test results) and when it should be recorded.

- Procedures for documentation, i.e., identifying and maintaining test record files including crew and scenario details, data collected, and test conductor logs.

(2) Where possible the use of a double-blind procedure should be used to minimize the opportunity of tester expectancy bias or participant response bias.

2.7.3 Test Conductor Training

(1) Test conductor personnel should receive training on:

- The use and importance of test procedures
- Experimenter bias and the types of errors that may be introduced into test data through the failure of test conductors to accurately follow test procedures or interact properly with participants
- The importance of accurately documenting problems that arise in the course of testing, even if due to test conductor oversight or error.

2.7.4 Participant Training

(1) Participant training should be of high fidelity; i.e., highly similar to that which plant personnel will receive in an actual plant. The participants should be trained to ensure that their knowledge of concept of the operator's role, concept of operations, the plant design, and use of the HSI is representative of anticipated users of the plant. It may be possible to limit training to the scope of the validation tests, however, participants should not be trained specifically to perform the validation scenarios.

(2) Participants should be trained to near asymptotic performance and tested prior to conducting actual test trials. Performance criteria should be similar to that which will be applied to actual plant personnel.

2.7.5 Pilot Testing

(1) A pilot study should be conducted prior to conducting the integrated validation tests to provide an opportunity to assess the adequacy of the test design, performance measures, and data collection methods.

(2) If possible, participants who will operate the integrated system in the validation tests should not be used in the pilot study. If the pilot study must be conducted using the validation test participants, then:

- The scenarios used for the pilot study should be different from those used in the validation tests, and
- Care should be given to ensure that the participants do not become so familiar with the data collection process that it may result in response bias.

2.8 Data Analysis and Interpretation

(1) Validation test data should be analyzed through a combination of quantitative and qualitative methods. The relationship between observed performance data and the established performance criteria should be clearly established and justified based upon the analyses performed.

(2) For all performance measures, informative descriptive statistics such as measures of central tendency and variability should be provided and compared to performance criteria. More rigorous analysis of data should be performed where possible.

(3) The degree of convergence of the multiple measures of performance should be evaluated.

(4) The data analyses should be independently verified for correctness.

(5) The inference from observed performance to estimated real-world performance should allow for margin of error.

(6) All design deficiencies should be corrected before validation efforts are concluded. Where it is not possible to fully correct a deficiency, justification should be provided and a alternative resolution of the human performance issue should be identified.

2.9 Validation Conclusions

(1) The statistical and logical basis for the determination that performance of the integrated system is and will be acceptable should be clearly documented.

(2) Final validation conclusions should include a consideration of the possible threats to:

- System representation validity
  - Inadequate process/plant model fidelity
  - Inadequate HSI fidelity
  - Inadequate participant fidelity
  - Participant sampling bias
  - Historical population changes
  - Operational conditions sampling bias
  - Inadequate scenario fidelity
- Performance representation validity
  - Test-level underspecification
  - Measurement underspecification
  - Changing measures
  - Poor measurement characteristics
  - Underspecified performance criteria
  - Measurement-scenario interaction.
- Test design validity
  - Test procedure underspecification bias
  - Tester expectancy bias

- Participant response bias
- Test environment bias
- Changes in participants over time
- Participant assignment bias
- Sequence effects

• Statistical conclusion validity

- Accepting narrow performance margins (difference between observed performance and a criterion) as acceptable (this may be due to an incorrect null hypothesis or inadequate consideration of performance variation)
- Low sample size
- High noise in data

(3) Validation limitations should be considered in terms of identifying their possible effects on validation conclusions and impact on design implementation. These should include:

• Threats to validity that were not well controlled

• Potential differences between the test situation and actual operations, such as absence of productivity-safety conflicts

• Potential differences between the validated design and plant as built (if validation is directed to an actual plant under construction where such information is available or a new design using validation results of a predecessor).

Part II

Criteria Development

and Technical Basis

3. DEVELOPMENT METHODOLOGY

The importance of complex human-machine system validation is widely recognized in the general systems development literature, by professional standards development groups, system designers, and by authorities who regulate such systems. However, there are few published standards or guidance documents available that provide methodological details and review criteria to support validation efforts, although there are numerous current efforts to do so. The documents that are available are predominantly scoping in nature, i.e., identify the scope of validation but treat its methodological aspects at a very general level. Thus, Meister (1986) noted that the literature on system test and evaluation is slim in comparison to the literature on analysis. The lack of guidance on appropriate integrated system methods has been noted by others as well (Wise, et al., 1994). This need has given rise to several recent efforts to improve the technical basis upon which validation methods can be developed. For example, the North Atlantic Treaty Organization (NATO) has sponsored a recent symposium devoted to human factors engineering (HFE) validation of complex systems (Wise, Hopkin, and Stager, 1993) and a standards development effort by the International Electrotechnical Commission (IEC, in preparation) has been initiated to provide a verification and V&V standard for the nuclear power industry.

The lack of validation guidance was noted in the development of the HFE PRM as well. Thus, this project was conducted with the objective of developing more comprehensive criteria for the review of integrated system validation.

A general approach to validation was developed as the first step to review criteria development. This general approach is referred to as a validation paradigm. A paradigm is defined as an example serving as a model or pattern. The paradigm provides a conceptual approach to validation by identifying important validation principles and their relationships. The general concepts in the paradigm are concerned with (1) establishing the requirements for making a logical and defensible inferences from validation tests to predicted integrated system performance under actual operational conditions, and (2) identifying the aspects of validation methodology that are important to the inference process. While it is recognized that differences in specific methodologies are possible, the general principles and concepts that are described by the paradigm are invariant across methodologies. The integrated system validation paradigm was developed using (1) the existing HFE PRM review criteria; (2) system test, evaluation, and validation literature; and (3) principles adopted from scientific research methodology.

A broad base of validation literature was reviewed. In addition to those standards and guidelines addressing validation, research and engineering literature was used to identify the current state-of-the-art in validation concepts. The literature review was focused on the validation of complex human-machine systems (see Section 4.1 below), in contrast to evaluations of more limited systems (such as software usability tests), or development tests (such as prototype evaluations). The scope of the literature review included current standards and guidance documents on validation related to complex system such as those found in the nuclear and defense industries.

To augment the published standards and guidelines addressing validation, principles were adopted from human performance research methodology. While important differences between validation and research are recognized, the logic required for validation and research have many

similarities with respect to the decision and inference processes required. The aspects of the process common to both endeavors include:

- Hypotheses are developed,
- Conditions relevant to the hypotheses are identified,
- Performance measures are defined,
- Controlled observations (measurements) are obtained under relevant conditions,
- Data is analyzed to examine the hypotheses,
- Conclusions are drawn with regard to the hypotheses, and
- Considerations of the generalizability of the results are made.

In the development of scientific knowledge about human performance, an hypothesis is identified as a logical deduction from theory. The hypothesis specifies the predicted relationship for an independent variable(s) and a dependent variable(s). Next , an experiment that provides suitable test conditions is developed to allow data to be collected that can be used to test the hypothesis. The data is then analyzed to make estimates about the characteristics (parameters) of the population that was sampled. Generalization of the results is based upon an inference process that considers:

- The quality of the experimental methodology (e.g., freedom from confounds),

- The quality of the measurement process,

- The statistical basis for the generalization (i.e., the probability that the data observed are the result of chance rather than random error or variability), and

- The degree to which the experimental variables were representative of the way the same factors are characterized in the population to which the results are to be generalized.

Validation is not a test of theoretically derived hypotheses or a formal experiment, as described above. However, there are important parallels in the logic required. In integrated system validation, it is hypothesized that the system will perform acceptably based upon performance requirements that are developed using engineering analyses, experience, and judgement. While in scientific research one is typically interested in the effects of one or a small number of independent variables on an small number of dependent variables, the considerations in validation are primarily on establishing that observed performance meets performance requirements when the integrated system is subjected to the effects of a broad combination of independent variables or conditions that are expected contribute to variation in the system's performance. Instead of attempting to isolate the effects of one independent variable, while holding all others constant (or controlled), validation seeks to establish that performance under the variation of all important independent variables is acceptable. In research, one is typically interested in the relative relationships between independent variables and dependent variables. In validation, rather than focus on relative relationships, one is typically interested in establishing that specific performance criteria are met. Only when performance criteria are not met, is one interested in examining specific conditions to determine which led to unacceptable performance. A summary of some of the differences between validation and research are provided in Table 2.1.

It is important to point out that not all scientific research take place under highly-controlled, laboratory conditions. Research and evaluation methodologies have been developed to address applied issues and field settings, where the researcher cannot use rigorous experimental controls. These methods are valuable in developing an approach to validation because such methods require greater attention to the problem of causal inference. The methods developed to address less controlled settings are referred to as "quasi-experimental" (Cook and Campbell, 1979). Integrated system validation is very much like a quasi experiment, therefore, quasi-experimental principles formed a major technical basis upon which our approach to validation and the associated review criteria were developed.

Thus, while differences do exist between the characteristics of research and validation, the general research methodology, especially that associated with quasi-experimentation, involves many important concepts that are valuable to the development of a validation paradigm and its methodology. Research principles, concepts, and methods provide information which, when integrated with the HFE PRM and existing validation literature, form a solid scientific and technical basis for the development of a validation paradigm. The analysis of research methodology is also valuable in that it helps fill in the gaps in existing validation methodology.

Once the validation paradigm was identified, considerations were made as to the methodological aspects of the validation process that are needed to meet the general paradigm requirements. That is, while the paradigm identifies the requirements of the inference process, the next task was to identify a means by which the paradigm requirements can be satisfied. Based upon the detailed methodological considerations, criteria were then developed that would enable one to review either an HFE integrated system validation plan or the results of an actual validation program. The criteria also will allow one to identify any weaknesses or threats to the inference process that is necessary for the validation.

Table 3.1 A Comparison of Research and Validation Characteristics

DIMENSION	RESEARCH PARADIGM	VALIDATION PARADIGM
Purpose	To develop human performance theory	To evaluate integrated system performance
Objectives	To test causal relationships; i.e., test theoretically-derived hypotheses regarding the effects of independent variables on dependent variables	Determine if the integrated system performance is within identified requirements/criteria
Independent Variables	Tests of hypotheses involve relatively few independent variables	Tests will involve a relatively large set of "independent variables" to assure that all variables that are expected to have a significant impact on system performance are represented
Dependent Variables	Relatively few are selected to represent the aspect of system/human performance specified by the hypothesis	Many are selected to provide a comprehensive hierarchal evaluation of personnel-system performance
Participants	Requirements vary based on the nature of the hypotheses and populations to which results will be generalized	Participants are highly qualified and trained personnel who are representative of the user population
Scenarios	Scenarios are designed to accentuate differences in performance between levels of the independent variables (maximize primary variance)	Scenarios are designed to represent a broad range of conditions that are feasible for system operation
Testbeds	Requirements vary based on the nature of the hypotheses	Testbeds are high-fidelity representations of the HSI and underlying process
Statistical Models	The focus is on comparisons of relative performance	The focus is on comparisons to established performance criteria
Null Hypothesis	There is no significant difference between levels of the independent variables(s) or their interactions	Integrated system performance is not acceptable
Statistical Inference	Inferences are made to population parameters	Inferences are made to predicted ranges of system performance
Generalization	Generalization is usually a secondary consideration	The ability to predict performance of the actual system, based on observed data, is the primary consideration in validation

4. GENERAL VALIDATION PARADIGM

4.1 General Concepts

4.1.1 The Issue of Complexity

A complex human-machine system may be defined as one supporting a dynamic process involving a large number of elements that interact in many different ways. Some important characteristics of systems exhibiting such complexity include: close physical proximity of elements, common-mode connections, interconnectedness of subsystems, feedback loops, multiple and interacting controls, and indirect information (Perrow, 1984; Rasmussen, 1988). In addition to interactive complexity, another characteristic of complex systems is tight coupling. Tight coupling of systems is characteristic of time-dependent processes. The success of the process is dependent on precise changes in multiple subsystems which affect each other. The process is invariant, there is basically only one way for it to function in order to successfully achieve its mission. Deviations in parts of the system result in the entire system deviating from its proper functioning. Due to the characteristics of tight coupling, safety is addressed through preplanning; i.e., designers consider the types of failures that are most likely to occur and those of high consequence, and design their solutions in advance.

Modern NPPs are highly-automated, complex systems whose performance is the result of an intricate interaction of human and system control. This interaction creates a great opportunity for variability in overall plant behavior in response to events. A difficulty of complex systems is that they fail in complex ways often unanticipated by the designer. Events which are unanticipated by designers and unfamiliar to plant personnel pose the greatest threat to nuclear power plant safety (Vincente, 1992).

These characteristics of complex systems have implications for the personnel who are responsible for system operation and maintenance. Control of a complex system poses demands on the cognitive capabilities of the operators, both individually and as a crew (Woods et al., 1994). Reason (1987, 1988, 1990) refers to such a system as a complex multiple-dynamic configuration; that is, a problem-solving environment where the system changes as a result of the operator's actions and automatic control processes. The interaction of personnel and control system actions upon the plant sometimes creates variability in overall plant behavior that is not easily understood by plant personnel. This challenge may result in a reliance on decision-making heuristics which can increase the probability of human error. Interactions not anticipated may represent a form of "resident pathogen" (Reason, 1990), which are latent until the right set of circumstances trigger them.

Woods et al. (1994) analyzed numerous incidents involving complex systems and identified several common factors:

- The situations evolve from numerous failures rather than one large failure.
- Some of the factors are latent.
- The contributing factors include both personnel and system elements.

There are additional factors that increase the cognitive demands in complex systems. A few of these factors will be briefly discussed below, including: inferential and hierarchal processes, pace of dynamics, redundancy and reliability, and conflicting objectives.

Higher-level functions depend on plant processes, which are dependent on plant systems, which are in turn dependent on system components. Personnel intervention can occur at different levels in the hierarchy. Because NPP operators cannot observe the process directly, they must infer performance from a myriad of indicators, which provide information about various aspects of performance. Complex system performance is a property that emerges from the integration of all the components; it is not simply a product of them. As a result, it may be difficult to predict the performance of the integrated system based on component properties (Rosness, 1993; Wieringa and Stassen, 1993).

Monitoring and control by personnel may be more difficult in situations where events move at a pace slower than that for which clear feedback can be obtained on the effects of operator or automatic actions. "System lags in general are harmful to performance" (Wickens, 1986). In complex systems, such as NPPs, there are numerous sources of time lag, including the dynamics of the process itself and the characteristics of the HSI, which make it difficult for operators to evaluate the results of their actions. These characteristics are made worse when process disturbances slowly evolve through the occurrence of numerous small human and machine failures, as is typically the case with significant incidents at nuclear plants (Woods et al., 1994). Thus, slowly-evolving changes in plant states can make it difficult for plant personnel to maintain accurate situation awareness of plant conditions.

The redundancy and overall reliability of complex systems can make failures more difficult to detect. When failures actually do occur, operators often do not initially believe the validity of the information, instead assuming that alarms or indications stem from other problems such as miscalibrations or maintenance activities. This remains true despite much training emphasis to "believe your indications." Perhaps this stems from the fact that failures in indicators are more common than failures in the process. Also, with redundancy, failures in single components may be less visible to operators because redundant backup systems compensate for them.

The inappropriate response by operators to conflicting objectives plays a role in many accident situations. For NPP personnel, demands to maintain power production may conflict with demands to maintain safety. For example, operators may feel compelled to refrain from actions that will cause costly or long-term maintenance. At the same time, operators are responsible for plant safety. Situations arise when the trade-offs between these responsibilities are difficult to make in real-time. The incident at the Davis Besse plant (NRC, 1985) where an operator did not initiate decay heat removal using the feed and bleed method, although called for by procedures, is an example of this type of tradeoff. In this case the feed and bleed would have released radioactive water directly into containment necessitating extensive down-time for clean-up.

To illustrate the cognitive challenges of complex systems, Vincente (1991, p. 1) cites testimony of a Three-Mile Island operator, which illustrates several of these issues:

> "Let me make a statement about the indications. All you can say about them is that they are designed to provide indications of whatever anticipated casualties you might have. If you

> go…beyond what the designers think might happen, then the indications are insufficient and they lead you to make wrong inferences. In other words, what you are seeing on the gauge, like what I saw on the high pressurizer level, I thought it was due to excess inventory… I was interpreting the gage based on the emergency procedure, where the emergency procedure is based on the design casualties. So the indications then are based upon my interpretation. Hardly any of the measurements that we have are direct indications of what is going on in the system."*

One way to try to prevent accidents is to attempt to identify any deficiencies in system design prior to their emergence under actual system operation (Reason, 1990). Integrated system validation provides such an opportunity. However, new approaches to complex system validation are needed. Historically, complex systems have been "validated" when the reliability and acceptability of their components have been demonstrated. However, since the interaction between components (hardware and software) and personnel is significant, component level approaches to evaluating the acceptability of complex systems are insufficient. That is, it cannot be assumed that the integrated system will achieve its objectives merely because all of the subsystems and components, in isolation, achieve theirs. Validation must evaluate the performance of all these subsystems and components. Similarly, Rasmussen (1988) has indicated that complex systems "cannot be considered to have practically isolated internal functions, well contained by system boundaries and, therefore, adequately described by classical engineering analysis. Accidents happen when system boundaries break down. In this case, the preconditions for formal, mathematical analyses of system function also break down and the formal methods are replaced by different methods for analysis of accidents based on causal representations" (p. 6).

Thus complex systems present many challenges to the personnel who must operate and maintain the system. In light of these challenges, the validation of the HFE of such systems must ensure that the design minimizes these challenges. Where evaluations of less complex systems focus on the usability of the user interface, integrated validation must address the adequacy of performance of the entire human-machine system including personnel and their interactions with both the system and each other. However, complex system validation must deal with the number of plausible operational conditions that result for all possible interactions of systems, components, and personnel.

---

*In a PWR, such as the TMI NPP, it is crucial to maintain the primary system water inventory in the reactor at a level above the top of the fuel, in order to ensure adequate core cooling. Pre-TMI PWRs measured this by inference. That is, they measured the level of water in the pressurizer (PZR) tank, which acts as both a surge volume and pressurizing system for the primary. Level in the PZR is in turn measured by a differential pressure cell attached between the PZR and a reference leg. During the accident at TMI, the power-operated relief valve (PORV) on the PZR became stuck in the open position. This drained water from the primary system overall and also caused the pressure to drop. As the pressure dropped to saturation conditions in the primary, water flashed to steam in the reactor vessel and in the loops forcing water into the PZR. This water then flowed out the PORV at the top of the PZR. This led to a condition of inadequate core cooling in the reactor vessel. This situation also resulted in the indication of a high PZR level. This indication was interpreted by operators as having too much water in the primary rather than the true case of too little. The actual water level in the reactor was not measured. After TMI, plants were required to provide unambiguous indication of inadequate core cooling, such as reactor vessel level. This has provided better indications, although measurements are still somewhat indirect. As an example, margin to saturation is measured by computation from various temperatures and pressures. Reactor vessel level is measured, for example, by temperature-compensated, differential pressure instruments or by a heated junction thermocouple system that monitors the temperature difference as one proceeds upward in the reactor vessel.

4.1.2 Purpose of Complex Human-Machine Systems Validation

The HFE PRM states that the purpose of integrated system validation is to provide evidence that the integrated system adequately supports plant personnel in the safe operation of the plant; i.e., that the integrated design remains within acceptable performance envelopes. Indications of adequate personnel support include:

- Personnel tasks can be accomplished within time and performance criteria.
- The HSI will support a high degree of operating crew "situation awareness."
- The plant design and allocation of functions will provide acceptable workload levels to ensure a balance between vigilance and operator overload.
- The operator interfaces will minimize operator error and will provide for error detection and recovery capability.

This approach to validation is consistent with other approaches to validation discussed for nuclear plants, as well as other complex systems. Considerations of approaches developed or being developed by the International Electrotechnical Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), and the Department of Defense (DoD).

The draft IEC V&V standard (IEC, in preparation) defines validation as a test to evaluate whether the interaction between HSI design and personnel can support performance of crew functions, including safe, reliable operations.

IEEE Standard 1023 (IEEE, 1988) provides general guidance for the incorporation of HFE into the design of NPPs. Section 6.1.1.17, Final Test and Evaluation, indicates that "A final test and evaluation of the integrated system, including the human operators and maintainers, should be conducted to verify that all previously determined HFE criteria and requirements are met and that functional requirements are satisfied" (p. 15).

With respect to the design of military systems, the DoD requires that systems be subject to "operational testing"; the purpose of which is to test system effectiveness; i.e., to determine whether personnel can operate the system design to achieve the systems mission. Specific objectives of the evaluation are to (1) demonstrate conformance of system, equipment and facility design to human engineering design criteria; (2) confirm compliance with performance requirements; (3) obtain quantitative measures of system performance which are a function of the human interaction with equipment; and (4) determine whether undesirable design or procedural features have been introduced (DoD, 1979).

Thus there is a consistent view on the general purpose of validation in the literature.

4.1.3 Relationship of Integrated System Validation to System Development

Integrated system validation, as defined in this document, is not intended as the activity whereby HSI subsystem design concerns and issues (such as the coding techniques used in the alarm

system) are explored and evaluated. Such considerations should be addressed in system development tests and evaluations, which have a different set of purposes including resolving design tradeoffs, comparing design options, and ensuring that specific subsystem requirements are met. These types of evaluations are addressed in HFE PRM Element 7, HSI Design review.

This distinction is not always made in the general literature. It is, however, fully consistent with the distinction made within DoD between development and operational testing (Meister 1986, 1989). Development tests typically occur in the early and middle stages of development as the system concept as its detailed implementation are being defined. Such tests are typically directed toward specific issues associated with individual subsystems and do not necessarily involve the actual system operators. On the other hand, operational testing is conceptually different. It is not an extension of the system development process. Instead, it is an evaluation of whether personnel can operate the system design to achieve its intended mission.

4.1.4 Validation and Validity

The different uses of the terms validation and validity are a potential source of confusion. The term validation is used in this document to describe a process by which a NPP design is evaluated to determine whether it adequately satisfies the demands of the real-world operating environment. The term validity is used to describe characteristics of the methods and tools used in the validation process. Various forms of validity are discussed in Section 4.2.

4.2 Predicting System Performance: Validity and Inference

4.2.1 General Approach

While the purpose of validation is straightforward, how the validity of a design is established is not. Logically, a design can never truly be validated. The validation process cannot prove that a design meets all established criteria and will always perform acceptably under all real world circumstances. Such a proposition can never be logically defensible. Just as theory can never be proven (Popper, 1959), a design can only withstand challenge of being invalidated. Insofar as the design meets such a challenge, it is said to be validated. Therefore, validation principally establishes that, through a comprehensive validation evaluation, the design was not *invalidated*.

An important aspect of integrated system validation is the consideration of what constitutes a challenge to the operation of the system. Numerous views have been expressed in the literature that bare on this general question. Woods et al. (1994) emphasized that "Credible evaluations of human performance must be able to account for all of the complexity that confronts practitioners (personnel) and the strategies they adopt to cope with that complexity" (p. 102). Gould (1988) emphasized that an important part of system testing is the "try to find bugs, crash it, break it, etc." and that such efforts are of "immense value" in system evaluation (p. 772). Thus, an important aspect of validation is to detect design errors before they become lessons learned (Woods and Sarter, 1993). Rasmussen (1988) expressed a similar view that validation is the analysis of future conditions of use. Since complex systems fail in complex and often unanticipated ways, the validation process should significantly challenge the design and establish that performance of the actual system can reasonably be predicted to be acceptable under a broad range of plausible conditions. If performance is conceptualized in terms of statistical variance, then all significant variance components should be included in the equation.

Thus validation requires more than simply establishing that the integrated design can perform within acceptable performance envelopes. Such a finding is a necessary but not a sufficient condition for validating a design. A validated design is one that was tested using an evaluation methodology which provides a logically acceptable basis to predict plant performance on the basis of observed samples of test data. The validation process must specify what evidence is necessary to validate the system. This includes consideration of the types of inferences that must be made from validation test results to predict that actual system performance will be acceptable.

For complex human-machine systems where failure can be a safety concern, testing actual systems under accident conditions in a real-world environment is not feasible or practical. Thus the tests have to be conducted using a testbed that is representative of the actual system. The preferred testbed is a full-mission, high-fidelity simulator with real-time, realistic performance dynamics. Such a configuration provides a context that approximates the real-world system. It allows performance of the fully integrated system to be observed without the potential dangers and costs inherent in the operation of actual systems in challenging situations such as equipment failures and accidents. In fact, if evaluations were limited to real systems, they would be necessarily inadequate because they could not address important safety aspects of system performance (Hollnagel, 1993).

Since it is not possible to identify and test all possible threats to a complex system, the test data, instead, represent performance of observed samples of integrated system performance. The final step in the logical inference chain requires generalization from the simulation evaluation to the performance under real-world conditions. This step completes the inferential process. Even though we may be reasonably confident that the validation test results could be repeated on the simulator, it does not necessarily logically follow that the results could be repeated in the real world.

A basis for generalizability emerges from the comparability of the psychological and physical processes of the test and actual situations (Kantowitz, 1992). When actual data comparisons can be made, the prediction of plant performance based on validation can be justified. However, except for a limited set of normal operational events, such a data comparison cannot be made. For integrated system validation, generalizability must be logically established. The goal is to achieve generalizability of test results which is accomplished by providing a level of realism in the test environment, i.e., system, participants, and test conditions, is representative of the environment to which the results are to be generalized. Generalizability is only supported when all conditions for valid inference are satisfied. An overview of these conditions are briefly described below and are elaborated in the following sections.

Thus, validating complex human-machine systems is a process that requires a carefully developed methodology that will permit:

- Collecting data on system performance in a *simulated* environment,
- *Sampling* the possible conditions, and
- Providing a defensible technical basis upon which to predict real-world performance across a broad range of conditions.

Since validation consists of an inference process, there are several possible outcomes that can occur when decisions about system acceptability are made based on inferences from test data to actual performance (see Table 4.1). Validation tests can:

- Correctly predict the acceptability of real-world performance.
- Incorrectly predict that the design is acceptable when actual performance is (or will be) unacceptable. This may be referred to as a "Type 1" error; i.e., incorrectly validating the design.
- Incorrectly predict that the design is unacceptable when actual performance is acceptable. This may be referred to as a "Type 2" error, incorrectly rejecting the design.
- Correctly predict the unacceptability of real-world performance.

Table 4.1 Validation Inference Decision Matrix

VALIDATION CONCLUSIONS	ACTUAL PLANT PERFORMANCE	
	Acceptable	**Not Acceptable**
Performance Acceptable (Design Validated)	Decision Correct	Type 1 Error
Performance Not Acceptable (Design Not Validated)	Type 2 Error	Decision Correct

There can be many reasons for making Type 1 and 2 decision errors. Generally they can be traced to failures in the inference process with respect to general requirements for valid causal inference. Four general forms of validity that are important to causal inference have been discussed in the research literature: external validity, construct validity, internal validity, and statistical conclusion validity (Cook and Campbell, 1979). Causal inference is undermined by factors that weaken any of these aspects to validity.

We have adapted these concepts by tailoring them to the specific objectives of integrated system validation and to accommodate the differences between research and validation methodology (as discussed in Section 3). Since the concepts have been adapted, the names of several have been modified to better reflect their meanings with respect to their more restricted application to integrated system validation. Thus, the following terms are used to define the forms of validity that are essential to integrated system validation:

- System representation validity (external validity),
- Performance representation validity (construct validity),
- Test design validity (internal validity), and
- Statistical conclusion validity.

The meaning and importance of each form of validity is briefly discussed below. The aspect of the validation process that impact each type of validity are identified in general terms. (The validation methodology is discussed in greater detail in Section 5.) Also discussed are the threats to each type of validity. Threats to any one component of validity threatens the ability to make inferences from validation tests to the prediction of actual plant performance and can lead to the types of decision errors illustrated in Table 4.1.

The general questions addressed by each form of validity are summarized in Figure 4.1. Also identified in the figure are the types of methodological considerations that are important to supporting valid inferences.

4.2.2 System Representation Validity

Definition of System Representation Validity

System representation validity refers to the degree to which the integrated system validation tests include those aspects of the integrated system that are important to real-world conditions. Specifically, system representation validity is based on the representativeness of the:

- Process/plant model
- Human-system interface (HSI)
- Personnel
- Operational conditions.

Figure 4.1 Validity of Inference to Actual System Performance

Important Components of System Representation Validity

System representation validity is supported to the extent that each aspect of the system is representative of the actual system and its operation. When considering representativeness, one must consider which aspects of the integrated system are constant and which are variable. Constants aspects do not change; i.e., are not variable. The process/plant model and the HSI are constants because they have well defined characteristics that are always present during test scenarios (note, this does not mean they are not dynamic). HSI is broadly defined to include procedures, job support aids, etc. For constant aspects of the system, representativeness is supported by physical and functional fidelity. The higher the fidelity, the greater the model and HSI are representative of the actual plant. High fidelity is required because human performance is greatly affected by the detailed design characteristics. Patrick (1987) stated that "even slight changes in both the nature of the information available and the manner in which it is represented might have serious effects on performance" (p. 341). Note that if validation could be performed using the actual plant, they would be fully representative of the design to be validated. The degree to which the model and HSI deviate from the actual design, representativeness is compromised.

Personnel and operational conditions are variable aspects of the integrated system. Variable aspects do change and, thus cannot be completely represented. The entire population of possible operators cannot be included in validation tests, nor can the entire population of possible operational events. Two aspects of variable components of the system need to be considered: fidelity and sampling. The meaning of fidelity is similar to that discussed above for constants. The representation of personnel and scenarios should be as close to the actual plant as possible.

With regard to sampling, consideration should be given to attributes or characteristics of personnel and operational events that can reasonably be expected to cause variation in integrated system performance. Those which are expected to contribute to system performance variation should be specifically identified and a sampling process should ensure that variation along important dimensions of these attributes/characteristics is included in the validation tests. For example, level of experience is a personnel factor that can be expected to contribute to personnel and ultimately plant performance variability. Therefore, in sampling of personnel to participate in validation tests, variability in levels of experience should be included in the selecting participants. As another example, the degree to which operator actions are proceduralized, i.e., whether explicit detailed procedures are available, can be expected to contribute to performance variability. Therefore, in sampling of operational conditions for inclusion in the validation tests, variability in levels of proceduralization should be included; i.e., some scenarios in which operator action is guided by procedures and some where operator actions are not well defined by available procedures.

In the generalization of validation tests to actual performance, system representation validity is supported to the extent that: (1) important aspects of the integrated system are represented with high fidelity, and (2) important contributors to potential system performance variability have been included in the validation process.

Major Threats to System Representation Validity

1. Inadequate process/plant model fidelity - This is a threat to validity posed by inadequate fidelity of the model, e.g., inability to accurately simulate important functions, processes, systems, components and their interactions; an ability to provide information that accurately represents the behavior of the reference plant; or inaccurate time dynamics of the process or interaction between HSI and plant components.

2. Inadequate HSI fidelity - This is a threat to validity posed by inadequate fidelity of the HSI, e.g., inaccurate functional characteristics of the HSI; inaccurate or missing HSI components; or inaccurate physical representation of HSI components.

3. Inadequate participant fidelity - This is a threat to validity posed by the use of participants not from the population to which results are to be generalized, e.g., use of engineers or instructors as plant operators.

4. Participant sampling bias - This is a threat to validity posed by inadequate sampling of the relevant participant characteristics expected to cause variability in system performance, e.g., use of senior plant operators only.

5. Historical population changes - This is a threat that may arise due to changes in the characteristics of the population to which results are to be generalized which occur after the validation sampling process. Even if the original sample fidelity and sampling were completely adequate, consideration should be given to possible significant changes that may have occurred in the target population, such as changes in operator qualification requirements. In such a case, the original sample may be no longer representative of the current population.

6. Operational conditions sampling bias - This is a threat to validity posed by inadequate sampling of operational conditions such that not all significant demands imposed by characteristics of operational events to variability in system performance are included in validation tests. Limiting tests to design basis accidents only would be an example of this validity threat.

7. Inadequate scenario fidelity - This is a threat to validity posed by a failure to represent in validation scenarios those aspects of operational conditions that have a significant affect on human performance, e.g., the use of oversimplified scenarios.

4.2.3 Performance Representation Validity

Definition of Performance Representation Validity

A concept such as safety of integrated system performance is multidimensional. Therefore, many different variables can be selected to measure it. Performance representation validity refers to the degree to which performance measures adequately represent those performance characteristics that are important to safety. Thus performance representation validity is supported when a measure is representative of the aspect of performance to be measured. For example, using the subjective opinion of operators as the measure of plant operability would be inadequate as the sole measure of safety.

While such opinions may be an important aspect to the overall evaluation, they are not necessarily predictive of actual performance of the plant.

Important Components of Performance Representation Validity

There are two aspects to performance representation validity that need to be considered: performance measurement selection and criteria definition. When validating integrated system performance, plant safety is directly indicated by the performance of the plant functions and systems that is most directly tied to safety. Thus, it may seem appropriate to consider only such measures as the validation criteria, i.e., if the plant remains within function and system performance criteria, the design is validated. However, such an approach is not sufficient for integrated system validation. Measures of human performance must also be included. Reasons for including such measures are described below.

Operators contribute to the plant's defense-in-depth approach to safety. The defense-in-depth approach to NPP safety is challenged if one of its main components is not performing at acceptable levels, e.g., that a safety system was functioning at capacity with no performance margin. Personnel serve a vital function for control of the plant. Therefore, personnel performance should be measured.

Plant performance measures do not adequately describe human performance. Such measures have frequently been found to be insensitive to effects of the design on human performance parameters (Gartner and Murphy, 1976; Hart and Wickens, 1990; Meshkati and Lowenthal, 1988; Williges and Wierwille, 1979). The skill and expertise of highly trained operators can often compensate for inadequate design, however, there may be significant costs to personnel, such as poor situation awareness, high workload, and high stress. While professional operators can perform acceptably under conditions of very high workload (Bittner, 1992), such a situation is not acceptable since under real-world conditions and where such conditions may remain for sustained periods of time. At worst human performance can fail leading to potential plant safety problems, and at best, can lead to reduced performance margin.

In addition, plant performance measures may not provide adequate information to indicate the cause of inadequate performance. If poor plant performance is observed due to human failures one should determine the causes, e.g., failure to follow procedures, misdiagnosis, and high workload. Performance measures at the plant level will not provide a basis for conducting such root cause determinations. Measures that go beyond plant-level measures are needed to help identify the causes of inadequate plant performance.

A more comprehensive approach to evaluation is necessary to adequately assess important aspects of performance (Meister, 1986; Kantowitz, 1992; and Bittner, 1992). Deciding what aspects of an integrated system to measure beyond those reflecting plant performance is a significant validation consideration. Kantowitz (1990) has stated that "The fundamental problem of measurement is deciding what to measure. Theory can help answer this question be telling us where to look in complex system environments" (p. 1258). The important contribution of theory in the selection of performance measures is widely recognized (Vreuls and Obermayer, 1985). Consideration of NPP validation from such a perspective can help identify the important aspects of personnel and plant performance.

The operator's role in a NPP is that of a supervisory controller, i.e, plant performance is the result of the interaction of human and automatic control. Reason (1990) called this a complex multiple-dynamic configuration which is a difficult one for personnel to handle when things go wrong. Figure 4.2 presents a simplified representation of such a system. In addition to plant process failures, the automatic control systems and HSI can also fail. Thus, personnel must respond to failures of the plant and to the interfaces that communicate plant failures. Further, plant personnel can exercise control over automatic systems to initiate or terminate their actions.

The operator's impact on plant safety is mediated by a causal chain from the operator's physiological and cognitive processes, to operator task performance, and ultimately to plant performance through the operator's manipulation of the plant's HSI. With respect to personnel, HSI design impacts plant performance through tasks which are accomplished in support of their role in plant operations. The accomplishment of their role can be conceptualized as involving two types of tasks. Primary tasks are those involved in performing the functional role of the operator to supervise the plant; i.e., process monitoring, decision-making, and control. Secondary tasks are those the operator must perform when interfacing with the plant, but which are not directed to the primary task, e.g., navigating through and paging displays, searching for data, choosing between multiple ways of accomplishing the same task, and deciding how to configure the interface.

Figure 4.2 Relationship of Personnel and Automatic Systems in Plant Performance

To adequately perform their tasks personnel need to have a reasonably accurate assessment of the plant conditions. How an operator oversees the process, makes decisions, and takes action is largely tied to cognitive processes involved in developing and maintaining situation awareness. The ability to maintain situation awareness is related to workload (operators should perform best when the workload level is moderate since low levels lead to boredom and high levels result on performance decrements)(Huey and Wickens, 1993).

Workload is an important aspect of human performance for several reasons. First, one HFE concern frequently associated with complex, computer-based systems is information and cognitive overload. Second, supervisory control systems tend to be characterized by periods of relative inactivity followed by periods of intense activity when systems fail. These workload transitions periods are critical from a human performance perspective because personnel are more susceptible to errors during them. Third, computer-based HSIs tend to impose a significant amount of secondary task workload associated with interface management. It is possible for plant and personnel task performance to be acceptable and situation awareness to be accurate, but for workload to be very high. When the cognitive processes needed to develop situation awareness and manage workload are challenged, poor task performance may result and plant performance may be compromised. Integrated system validation should identify such effects, which only become apparent when personnel tasks are performed in the full context of the integrated system.

The hierarchal relationship between performance measures are illustrated in Figure 4.3. The measures used for validation should be designed to reflect the determinants of plant performance at each level. The purpose of a comprehensive, hierarchal approach to performance measurement is to assure margin in performance since a failures at any level in the hierarchy can impact performance at levels above. TMI was a excellent example of the importance of lower levels, e.g., failures of situation assessment, on subsequent component and system failures and ultimately to loss of critical safety functions.

Performance representation validity is based on the specific measures used in validation tests. The measures must be representative of the category of performance being evaluated. In addition, each measure should have reasonably good measurement characteristics, such as reliability. Once validation performance measures are selected, performance criteria must be established. Performance criteria are the standards against which the observed integrated system performance is compared to judge its acceptability.

Section 5.6 provides a discussion of representative measures, measurement characteristics, and criteria for achieving high levels of performance representation validity.

Figure 4.3 Hierarchal Performance for a Supervisory Control System

Major Threats to Performance Representation Validity

1. Test-level underspecification - This is a problem characterized by inadequate comprehensiveness of the measures. Examples include measuring integrated system validation using only operator task performance or only plant performance.

2. Measurement underspecification - Some variables are appropriately quantified by taking several measures at the same time while others are appropriately measured over time. Reactor temperature is an example of the former and heart rate is an example of the latter. Taking a single measure of reactor temperature would represent measurement underspecification because temperatures vary across locations of the RCS and sensors at these locations have different measurement characteristics. Thus, the value of reactor vessel temperature should be derived for several independent measures. Taking a heart rate measure at one point in time would represent measurement underspecification because it ignores the dynamic changes in heart rate over time.

3. Changing measures - This is a problem characterized by variation in measurement collected over time that occurs because the specific measures or measuring instruments are different at different points in time during the tests. An example would be using one scale to measure workload during early test trials and a different scale at a later point in time.

4. Poor measurement characteristics - At the level of individual performance measures, the factors that can undermine performance representation validity are any that can result in poor measurement characteristics, such as poor reliability, intrusiveness, and lack of sensitivity. Considerations of the properties of individual measures is essential to the validation process. A list of these factors is provided in Section 5.6.1.

5. Underspecified performance criteria - A performance criterion is the standard against which the integrated plant performance is compared for a given performance measures. Since the specification of these criteria are dependent on engineering analyses and human performance assessments, flaws in these analyses will lead to incorrect or poorly supported criteria.

6. Measurement-scenario interactions - This refers to a changes in performance that occur because the measurement technique interacts with the test scenario. For example, questions that are posed to participants to measure situation awareness during a scenario may influence the performance of participants, e.g., by directing them to seek certain information that they would not have otherwise sought. Such effects can be distracting and interfere with performance or they can cue participants to relevant situations, parameters, and events.

4.2.4 Test Design Validity

Definition of Test Design Validity

Test design validity addresses those considerations that are involved in the actual conduct of the validation tests. It includes activities such as the assignment of crews to scenarios, development of test procedures, and participant training. Even when the validity of the integrated system and measures are supported, the way in which the tests are conducted can undermine the logical linkage of the integrated system and observed performance. That is, aspects of the test design can alter the relationship between the integrated system and observations of performance, and thereby limit the generalizability of the validation test results to actual plant performance. When factors are introduced by the test design which systematically corrupt the interpretation of the system-performance correlation, test design validity is compromised.

Important Components of Test Design Validity

The three most common problems are biasing, confounding, and masking. *Biasing* is an aspect of the methodology which systematically modifies performance. For example, the instructions given to a operating crew prior to participating in a scenario could bias their behavior, such as telling them to be careful because the procedures may be misleading at some point in the scenario. Such comments may produce behavior that is unique to the instructions, rather than a natural response to the demands of scenario events.

Confounding is the systematic coupling of one aspect of the test with another aspect of the test or an extraneous variable. Confounding makes important relationships ambiguous. Confounding can occur between variables of interest such as crew experience and scenario difficulty; or between a variable of interest and an extraneous variable, such as time-of-day sequence in which tests are conducted. For example, suppose less experienced crews were always assigned to difficult scenarios while easy scenarios were always given to the most experienced crews. In this case crew experience and scenario difficulty are said to be confounded. If one then observes that the integrated system performance is poor in difficult scenarios, one would not know if the result was due to the scenario difficulty or to the fact that the crews were inexperienced. Confounding can also preclude the detection of an important interaction. For example, if the assignment were opposite and all performance were acceptable, the test design might have failed to reveal that less experienced crews cannot handle the more difficult scenarios.

In general, confounding and biasing effects provide alternative explanations of validation test results, and thus, make the test results ambiguous. Factors responsible for such effects should be controlled and minimized.

Masking is the addition of noise or error variance to performance data, which makes the results more difficult to interpret and the prediction of actual plant performance less certain. For example, inconsistent instructions to participants can increase the noise in the data. Masking effects, like all sources of noise, should be minimized.

Major Threats to Test Design Validity

1. Test procedure underspecification bias - This bias occurs when test procedures do not include clear and objective instructions to test conductors regarding how the test should be conducted. Test instructions should address procedural concerns such as:

   - how to brief the participants,
   - when to start and stop scenarios,
   - when and how to interact with participants during scenarios,
   - when and how to collect measures, and
   - if and when bias and/or noise may be introduced into the test.

2. Tester expectancy bias - This is a bias in which the collection of data is systematically influenced by the expectations of the testers. This is different from test procedure underspecification bias because the bias is the product of expectations rather poorly specified procedures. This can exhibit itself in many ways. Testers may, through the provision of subtle cues or communications, provide direction to participants. For example, if the test conductors were also the system designers, they may tend to evaluate the performance of participants in ways that reflect more favorably upon the design than others would.

3. Participant response bias - Response bias means that the data obtained in a test are influenced by the test design itself. It is not necessarily implied that response bias represents any deliberate attempt by the participants to be untruthful. Humans in a test situation naturally respond to the test environment. The test environment can influence participants in ways that have little to do with the objectives of the tests. Characteristics of the test environment to which the participants respond, independent of the objectives of the test itself, are called demand characteristics.

Response bias can occur in four ways. First, participants may wish to influence outcomes and be biased toward producing data that is consistent with the desired result. Second, participants may want to provide data that they think the test conductors want

to obtain. For example, if the test conductors are also the system designers, participants may be reluctant to criticize the design. Third, participants may try to figure out how performance should vary in differ

ent condi tions and then influe nce data to be consi stent with such differ ences . Forth , partic ipant s may want to excel becau se they know that they are being obser ved. This is the well know n Hawt horne effect . (Coo k and Cam

pbell, 1979)

4. Test environment bias - Integrated system validation takes place in a testing environment. The test environment may have limitations with respect to creating the operational environment of an actual plant. The somewhat artificial nature of the test environment can modify personnel behavior, for example, with respect to (1) the influence of performance shaping factors (PSFs), and (2) important human information processing parameters. With respect to PSFs, simulator exercises will not reflect with high fidelity the influence of all important factors (such as stress, noise, chaos, distractions, and fatigue on late shifts) that will affect human performance during real-world operations. With respect to human information processing, important aspects of human cognition and performance (such as signal detection threshold, event probability estimation, and response selection) are affected by the operating crew's understanding that it is participating in a simulated rather than a real situation. For example, when a simulator exercise begins, operators know something other than normal operations are likely. Unlike the real world, very low probability events are likely to occur and will be anticipated by the crew. Thus, the operator's attention is aroused and focused on event occurrence and detection.

When a situation does occur, the crew's response will likely be optimized according to established procedures, because there are no real consequences to responses made on a simulator and no conflict between safety and produ

ctivit
y
(pow
er
produ
ction)
goals
.
How
ever,
in
real-
world
condi
tions,
there
are
major
conse
quen
ces to
actio
ns
and
these
conse
quen
ces
may
affect
the
proba
bility
and
timin
g of
opera
tor
actio
ns.
Conv
ersel
y,
opera
tors
may
not

take some risks that they would in the real world because their performance is being observed. All of these factors require the recognition of uncertainties in the use of simulator data. Good validation test procedures can help reduc

e this problem, but it cannot be completely eliminated. Therefore, the interpretation of results from simulator studies contains uncertainty that limits the generalizability of the results to real-world conditions. When possible, behavior

observed in simulator studies should be compared to behavior in actual systems.

5. Changes in participants over time - Participants are going to exhibit changes over the course of the validation testing due to numerous effects such as learning more about the HSI, becoming more familiar with the testing environment, and fatigue. Efforts should be made to limit these effects, for example, by providing prevalidation training to an acceptable performance criterion. However, participant changes cannot be totally eliminated. Therefore, tests should be arranged to balance such effects over test scenarios and thereby assure that the effects are not systematically confounded with conditions of interest (see Section 5.7.1, for a detailed discussion of this topic).

6. Participant assignment bias - This is a systematic bias in the assignment of test conditions to participants. The example above, regarding participant experience, is an example of this type of confound.

7. Sequence effects - This is a systematic bias due to the sequence in which scenarios are presented. If crews participate in more than one test scenario, the sequence of scenarios should be varied from crew to crew to avoid having a consistent influence of one scenario on the next.

4.2.5 Statistical Conclusion Validity

Definition of Statistical Conclusion Validity

Statistical conclusion validity addresses the relationship between the performance data and the established performance criteria. This relationship is not straightforward as will be discussed below. The fact that the observed performance is within an acceptable range, is a necessary but insufficient basis for establishing statistical conclusion validity. This is because the observed data represent only a sample from the population of performance data. It is the generalization to the population of performance that is of primary interest.

Important Components of Statistical Conclusion Validity

Performance of a complex task will vary, not only from crew to crew, but for the same crew from one scenario to the next, even under similar conditions. Thus, it necessary to consider the possibility that observed performance was due only to chance and that a different result would be obtained if the tests were repeated. In addition, the scenarios used during the validation tests represent only a sample of all possible scenarios. As a result, it is necessary to consider what can be reasonably inferred, using the observed performance about the relationship between estimated *population* performance and the performance criterion. Figure 4.4 illustrates these relationships. In general, one is not interested that the system performs on average within the acceptable envelope of performance. Rather one is interested in determining that no individual "trial" falls outside the acceptable performance envelope (the range of predicted performance).

Figure 4.4 Performance Range Relative to Performance Criterion

Data analysis is complicated by the fact that there may be no single strategy that is necessary to maintain performance within the acceptable bounds. Each crew may use a different strategy to maintain acceptable plant performance. Operators of complex systems do not always act to minimize the deviation of process parameters from an optimal value, mean, or function. Operators may, for example, exceed a prescribed rate of reactivity change in order to prevent a reactor trip. As a result, some descriptive statistics such as measures of central tendency, may be misleading because the individual crews deviate from the central value for different, yet acceptable, reasons.

Statistical conclusion validity can be understood with regard to the two types of decision errors that can be made, previously discussed in Section 4.2.1. It is important to point out that statistical conclusion validity is not, in and of itself, the complete decision basis for system validation. While statistical conclusion validity deals with the relationship between data and the performance criterion, final decisions regarding system validation require a consideration of each type of validity including how well the system was represented, how representative the performance measures were, and how well the test was conducted (see Section 4.3 below). With this bigger picture in mind, the discussion below addresses decision errors in the context of data analysis.

A "Type 1" error reflects an incorrect decision that the design is acceptable which results from favorable performance data that were obtained purely by chance; i.e., if the tests were repeated, the result would be unacceptable. Logically, the validation null hypothesis is that performance is unacceptable, therefore the burden of proof is to establish that the design is acceptable. Observed performance should lie significantly away from accepted performance bounds (margins) to conclude that population performance would be acceptable. The possibility of making this type of error cannot be totally eliminated, but it can be controlled. Statistical techniques are used in research to specify an acceptable probability of this type of error. Using this probability, called the significance level, a difference between conditions is considered reliable only if the probability of its having occurred by chance is less than the acceptable probability of error. For example, if the acceptable error rate is set at .01 (1 in 100) and the observed difference has a significance level below the rate (e.g., .001), then it would be considered statistically significant. That is, it would have less than an .05 (1 in 20) probability of being due to chance.

While the same logic may be applied to evaluating the results of validation data, a validation test may not yield sufficient data under constant conditions to permit statistical tests or analyses to be applied in the same way that one would in a research setting.

Another aspect of statistical conclusion validity is the overall sensitivity of the evaluation for detecting that performance would fall within acceptable limits. Concluding that the design is not acceptable when the actual performance would be acceptable is called a "Type 2" error. There can be many reasons for making this type of decision error. A limited data set, for example, generally increases the predicted range of performance. Poorly constructed test procedures may do so also because they may lead to inconsistencies in the way tests are conducted, which in turn, leads to increased noise or error variance. This is a problem of low statistical power. This is another reason to maintain a null hypothesis that performance is unacceptable. If, instead, the null hypothesis were that performance was acceptable, low power and test insensitivity would work in favor of validating the design. Such an approach would be unacceptable.

Another factor to consider with respect to statistical conclusion validity is the degree of convergence of the multiple measures of performance. When all the measures of performance are considered, there should be consistency of statistical conclusions.

Major Threats to Statistical Conclusion Validity

1. Accepting narrow performance margins - A performance margin is the difference between observed performance and a criterion. Accepting performance margins as acceptable which are two narrow can lead to a type 1 error. One way this error may occur is if the validation team

assumes an incorrect null hypothesis. That is, if the null hypothesis is that the design is acceptable unless demonstrated otherwise, then the design is considered acceptable if the observed performance lies anywhere within the acceptable region. With this null hypothesis, the design is considered unacceptable only when observed performance lies on the unacceptable side of the criterion. A second way this error can occur is when there is inadequate consideration of human performance variation and its generalization to actual performance. That is, observed performance is accepted as the exact measure of variation and there is no consideration that actual performance may vary more than that observed because of factors not foreseen or accounted for in the testing.

2. Low sample size - As a general rule, the larger the sample size (number of participating crews), the more confidence can be placed in generalizing the observed test performance to performance. Low sample sizes make it difficult to examine the effects of human variability. However, it should be recognized that there is a significant tradeoff between sample size and the difficulty, time, and cost of the validation program. Since human and integrated system variability is important to the generalization process, methods should be employed to ensure its adequate estimation.

3. High noise - Any aspect of the validation tests that lead to increased variation in test data constitute threats to statistical conclusion validity. For example, if test procedures do not include clear and objective instructions to test conductors, variability in test administration may result in differences in personnel performance that are not related to the design of the system. Poor training is another example of a source of noise in test data.

   A major source of noise can be poor measurement reliability. All measurements are a function of the true score plus error. The reliability of a measure is the estimation of the degree to which data

indic ate true score s. Since reliab ility (r) is typic ally expre ssed as a corre lation coeffi cient, 1-r provi des an indic ation of the degre e of error in the meas urem ent or noise . The highe r the value of r, the less the meas urem ent error

and the more confidence can be placed in the data obtained. The lower the value of r, the more error is associated with the measure and the data become difficult to interpret.

4.3 Characteristics of a Validated System

An integrated system is considered validated when the following conditions are satisfied:

1. A comprehensive testing program was conducted by an independent, multidisciplinary team.

2. System representation validity: The integrated system is representative of the actual system in all aspects that are important to integrated system performance. Constant aspects of the system

are high-fidelity and variable aspects of the system were adequately sampled and represented in high-fidelity.

3. Performance representation validity: The measures of integrated system performance and their associated criteria reflect good measurement practices and are concluded to be representative of important aspects of performance.

4. Statistical conclusion validity: Based upon a convergence of the multiple measures, it can be concluded that the performance of actual system will be acceptable.

5. Test procedure validity: There are no plausible biasing or confounding effects to make the predictions of system performance ambiguous.

When these conditions are met, the results of the validation process are considered acceptably representative of the actual system performance and generalization is supported. In essence, the validation test program has failed to invalidate the design.

4.4 Limits to the Predictability of Actual System Performance

There are clearly limits to which an integrated system validation can be expected to predict actual system performance. All predictions are made with some degree of decision error. One must recognize that:

- Not all the threats to validity can be completely eliminated or controlled.

- There may be subtle or even acknowledged differences between the validated design and the construction of the actual plant, which make the implemented design different from the integrated system that was validated.

- Integrated system validation will not typically include considerations or influences of organizational factors, such as safety culture and administrative procedure philosophy, which are important to the safe operation of the plant.

While limitations to integrated system validation are recognized, it is important to emphasize that the complete safety evaluation is based upon the establishment of convergent validatity (see Section 1.1). Thus, integrated system validation is one part of a comprehensive evaluation.

5. VALIDATION METHODOLOGY

This section describes the methodological considerations that should be addressed in the integrated system validation, including:

- Validation Team (Section 5.1)
- Test Objectives (Section 5.2)
- Validation Testbeds (Section 5.3)
- Plant Personnel (Section 5.4)

- Operational Conditions (Section 5.5)
- Performance Measurement (Section 5.6)
- Test Design (Section 5.7)
- Data Analysis and Interpretation (Section 5.8)
- Validation Conclusions (Section 5.9)

Each is discussed below.

5.1 Validation Team

A multidisciplinary team is needed to conduct an integrated system validation. Appropriate areas of expertise are described in Appendix A of the HFE PRM. Each of the technical disciplines listed in the HFE PRM may not be necessary. Rather, the specific technical areas of expertise required for the validation team should be based on the scope of the validation effort.
In addition, the validation team should include personnel with expertise in test and evaluation, including test design, test procedure development, performance measurement, and data analysis.

To support objectivity of the evaluation, the members of the validation team should have some degree of independence from the personnel responsible for the actual design. The purpose of an independent validation team is to help ensure an unbiased evaluation. The use of an independent validation team should avoid problems that occur when systems are tested against exactly the same constraints and assumptions to which they were designed. Independence also helps guard against tester expectancy bias as a threat to the validity of the validation process. Other NPP documents (EPRI, 1992; Draft IEC standard, in preparation) also support the concept of a validation team that is independent of the designers. (See also Woods and Sarter, 1993, for a discussion of the role of evaluator as a contributor to the design process.)

The team should have access to all HFE program documentation (e.g., design files, analyses, evaluations) and to the members of the HFE design team who were responsible for the development of design and documentation.

5.2 Test Objectives

The purpose of integrated system validation is to provide evidence that the integrated system adequately supports plant personnel performance in the safe operation of the plant. To accomplish this purpose, the validation test should address a full range of test objectives that relate to this purpose. Detailed objectives should be defined in a systematic manner which relates scenario characteristics and performance measurement criteria.

The general considerations that should be addressed in validation include:

- Validate the role of plant personnel, i.e., that the allocation of functions to human and automatic aspects of the integrated system are appropriate and takes advantage of human strengths and avoid allocating functions that would be negatively affected by human limitations.

- Validate that the shift staffing, assignment of tasks to crew members, and crew coordination (both within the control room as well as between the control room and local control stations and support centers) is acceptable. This should include validation of the nominal shift levels, minimal shift levels, and shift turnover.

- Validate that for each human function, the design provides adequate alerting, information, control, and feedback capability for human functions to be performed under normal plant evolutions, transients, design basis accidents, and selected, risk-significant events that are beyond-design basis.

- Validate that specific personnel tasks can be accomplished within time and performance criteria, with a high degree of operating crew situation awareness, and with acceptable workload levels that provide a balance between vigilance and operator burden. Validate that the operator interfaces minimize operator error and provide for error detection and recovery capability when errors occur.

- Validate that the functional requirements are met for the major HSI features, e.g., group-view display, alarm system, safety parameter display system (SPDS) function, general display system, procedures, controls, communication systems, controls EOP-related local control stations.

- Validate that the crew can make effective transitions between the HSI features in the accomplishment of their tasks and that interface management tasks such as display configuration and navigation are not a distraction or undue burden.

- Validate that the integrated system performance is tolerant of failures of individual HSI features.

- Identify aspects of the integrated system (including staffing, communications, and training) that may negatively impact integrated system performance.

More detailed objectives for each of the above general objectives should be developed by the applicant to reflect specific characteristics of the applicant's design.

5.3 Validation Testbeds

The HFE PRM states that integrated system validation should be performed by evaluating dynamic task performance using tools that are appropriate to the accomplishment of this objective. In Section 4.2.2, System Representation Validity of this report, the importance of representativeness of the process and plant model and the HSI with respect to the actual plant design was discussed. The degree to which the plant model and HSI deviate from the actual design determines the degree to which representativeness is compromised and, thus, the degree to which threats to system representation validity emerge. This section will consider the factors that should be considered in evaluation of the representativeness of the testbed.

Stubler, Roth, and Mumaw (1992) identified the dimensions of realism and completeness as significant considerations. Realism refers to the degree to which physical characteristics and functionality of the real HSI are included in the testbed (e.g., the representation of the human-machine system in the evaluation). Completeness refers to the degree to which the testbed represents the entire human-machine system.

Based upon consideration of testbeds in the literature, the aspects of the HSI and process model that are significant to integrated system validation are discussed below.

1. HSI completeness - refers to the degree to which the testbed represents the entire facility. A testbed may represent one aspect of the HSI, such as one panel or workstation, or the entire HSI, e.g., the main control room.

2. HSI physical fidelity - refers to the degree to which the physical characteristics of the actual plant HSI are included in the testbed. High physical fidelity in the HSI means that the HSI used for validation is essentially a replica in form, appearance, and layout of the design to be implemented in the actual plant.

3. HSI functional fidelity - refers to the degree to which the functional characteristics of the actual plant HSI are included in the testbed. This includes the way in which the HSI components operate, its modes of operation (e.g., the changes in functionality that can be invoked on the basis of operator selection and/or plant states), types of feedback provided, and its dynamic response characteristics (e.g., from data processing, the time required for actions such as time to draw displays or update parameter value).

4. Data completeness fidelity - refers to the degree to which the information and data presented at the HSI represents all plant processes, systems, and components. A testbed can represent the data associated with the entire plant or only a portion of the plant. For example, a testbed that is mainly concerned with use of the primary systems of a PWR plant may represent the secondary side of the plant with a low level of fidelity (e.g., provide only those parameters needed to support tasks performed on the primary side of the plant).

5. Data content fidelity - refers to the degree to which the HSI accurately presents the information and data associated with aspects of the plant modeled. The degree of fidelity is provided by the underlying plant model. High data content fidelity is supported when the information and controls presented at the HSI are based on an underlying model of the plant dynamics that accurately reflects the reference plant. Plant data presented by the testbed may be that of the actual plant, a high quality simulation of the plant, a different plant, a simplified model , or fictitious data.

6. Data dynamic fidelity - refers to the degree to which the changes in plant data presented by the testbed are depicted as they would in the actual plant. In validation tests, the dynamic response of plant information and data to changes in the plant is related to the process model. The dynamics of the plant response may be low fidelity such as in a static mock up or high fidelity such as a typical NPP training simulator.

7. Environment fidelity - refers to the degree to which environmental characteristics of the actual task environment that impact human performance are represented in the testbed. Considerations may include factors such as noise, lighting, heat, and ventilation. Environmental considerations are particularly important for tasks that are performed outside of a control room environment, such as a local control station where protective measures from environmental effects are needed; e.g., special clothing and equipment.

5.3.1 Representation of the Main Control Room

HSI Completeness

Meister (1986) has emphasized that to be ready for operational testing "the system must be a complete entity - no missing modules. Personnel to operate the test system must be representative of or similar to those who will eventually operate the system and must have been trained to do so.

Procedures to operate the system as it is designed to function operationally must have been written out" (p. 214). Similarly, the IEC draft V&V standard states that "the operator's activity is likely to be biased… unless a complete replica of the control room is available and tests are conducted with the involvement of the complete operating crew" (p. 34).

It may seem appropriate to provide an interface that includes only those HSI elements that are needed for the scenarios to be conducted as part of validation tests. While such part task or partial HSI representations may be sufficient for development tests, they are unacceptable for integrated system validation. Completeness is important in integrated system validation because it enables:

- The possibility for unwanted interactions to occur, e.g., between navigation features and control features of the interface.

- Human performance problems to be observed related to factors such as misuse of the HSI due to the distracting effects of aspects of the design not involved in the current task or delays in task execution due to search for the proper HSI element.

Partial HSI representations may preclude the potential for such human performance problems to be detected.

HSI Physical Fidelity

A high degree of physical fidelity in the HSI should be represented, including presentation of alarms, displays, controls, job aids, procedures, communications, interface management tools, layout and spatial relationships.

HSI Functional Fidelity

A high degree of functional fidelity in the HSI should be represented. All HSI functions should be available. High functional fidelity includes HSI component modes of operation, i.e., the changes in functionality that can be invoked on the basis of operator selection and/or plant states.

Data Completeness Fidelity

Information and data provided in the control room should completely represent the plant systems monitored and controlled from that facility.

Data Content Fidelity

A high degree of data content fidelity should be represented. The information and controls presented at the HSI should be based on an underlying model that accurately reflects the reference plant. The model should provide input to the HSI in a manner such that information accurately matches that which will be presented in the actual control room.

Data Dynamics Fidelity

A high degree of data dynamics fidelity should be represented. The process model should be capable of providing input to the HSI in a manner such that information flow and control responses

occur accurately and in a response time that matches that in the actual control room. Overall, the HSI should provide the same response times as the actual control room; e.g., information should be provided to the operator with the same delays as would occur in the plant. This is another reason for completeness, because the large amount of data to be processed in the final control room may result in added time delays.

Environment Fidelity

A high degree of environment fidelity should be represented. The lighting and noise characteristics of the control room should reasonably reflect that expected in the actual control room. Thus, in the design of validation tests, consideration should be given to the design of the control room lighting. The noise contributed by equipment, such as air handling units, and computers, etc. should be represented in validation tests.

One approach to providing a validation testbed that is consistent with the above fidelity discussion, is to use the American National Standard "Nuclear power plant simulators for use in operator training," (ANSI/ANS-3.5-1985) as a guide. This standard was found generally acceptable for achieving training simulator requirements by the NRC as described in Regulatory Guide 1.149, "Nuclear Power Plant Simulation Facilities for Use in Operator License Examinations (NRC, 1987a).

Within the framework of the standard, a full-scope simulator is defined as "A simulator incorporating detailed modeling of systems of the reference plant with which the operator interfaces in the control room environment. The control room operating consoles are included. Such a simulator demonstrates expected plant response to normal and off-normal conditions" (p.1). Several additional considerations include:

- The "reference plant" is the design being validated and not a similar or generic plant.

- The types of malfunctions should not necessarily be limited to those identified in Section 3.1.2, Plant Malfunctions, of the standard since more advanced plant designs may have functions, systems, components, and malfunction conditions that are different than those of conventional plants. The conditions developed using the sampling process described in Section 4.5, Operational Conditions, of this document should be used as a guide to the specific plant malfunctions necessary for integrated system validation.

- Section 3.2 of the standard, Simulator Environment, discusses the degree to which the representation of the HSI and environment may deviate from the reference plant. While some deviation is acceptable, validation tests should strive for greater fidelity than may be necessary to support training objectives. Any deviations should be justified in terms of why it could not be reasonably expected to impact performance.

- Section 3.4 of the standard, Simulator Training Capabilities, identifies a number of requirements that are specific to training objectives, e.g., the number of initialization conditions is 20. These considerations should be modified to meet the requirements of the validation test program. Section 5.5 of this report addresses the identification of operational conditions for evaluation. The validation test facility should be capable of providing initialization conditions for all scenarios constructed to meet the defined set of operational events.

- Section 3.4.4 of the standard, Instructor Interface, is not specifically needed. However, a test conductor's station is needed from which the simulations can be initialized, monitored, and terminated. From this location, the test conductors should be able to control equipment status, insert faults, control data collection, and act as surrogate personnel outside the control room as required by validation test scenarios.

- There may be additional requirements to support data collection as described in Section 5.6, Performance Measurement, of this document. Plant data should be available in computer form to facilitate data analysis.

A limited scope simulator as defined by the standard (i.e., a simulator incorporating limited modeling of a generic plant or subsystem design) would not be acceptable.

5.3.2 Representation of Monitoring and Control Facilities Remote from the Main Control Room

Validation tests may include scenarios where important actions are taken at remote shutdown facilities and at local control stations. They may also include important interactions between the main control room and support facilities such as the TSC or EOF. It may not be necessary to provide as high-level simulation of these facilities as for the main control room. The decisions to represent such facilities should consider importance of the actions to safety taken at the facility, complexity of the actions and the HSIs, and criticality to the overall test of the accuracy of the timing factors associated with personnel actions. For important actions at complex HSIs, where timely and precise human actions are required, the use of a simulation or mockup to verify that human performance requirements can be achieved should be considered.

When simulations or mockups are used, the important characteristics of the task-related HSIs and task environment (e.g., lighting, noise, heating and ventilation, and protective clothing and equipment) should be included in the testbed.

For less critical actions or where the HSI are not complex, e.g., a simple local panel where a key switch must be activated, it may be possible to represent the human performance based on analysis rather than simulation. For example, an analysis can be performed to determine a realistic time estimate for performing the action. This estimate should consider ANSI/ANS 58.8-1994, "American National Standard Time Response Design Criteria for Safety-Related Operator Actions." This estimate should also include factors such as communication time, time to go to the LCS, time to gain access (e.g., unlock panel doors), and time to gather information, make a decision, take action, and obtain feedback. These times should be used in the simulation to provide accurate response times.

5.3.3 Testbed Verification

Before a testbed is used for validation tests, it should be verified for conformance to the applicable seven aspects of testbeds discussed, i.e., HSI completeness, HSI physical fidelity, HSI functional fidelity, data completeness, data content fidelity, data dynamic fidelity, environment fidelity. Detailed design descriptions and documentation can be used as the basis for verification.

5.4 Plant Personnel

In Section 4.2.2, System Representation Validity, the importance of representativeness of personnel was discussed. Personnel, like operational conditions, is a variable aspect of the integrated system; i.e, an aspect of the system that changes. Thus, two aspects of a variable component need to be considered: fidelity and sampling. The meaning of fidelity is similar to that used in the discussion above of the HSI and process model. With regard to sampling, consideration should be given to attributes or characteristics of personnel that can reasonably be expected to cause variation in integrated system performance. Those that are expected to contribute to system performance variation should be specifically identified and a sampling process should ensure that variation along that dimension is included in the validation. To the extent that the sampling process is appropriately conducted, representativeness of personnel is supported. When representativeness is compromised, threats to system representation validity emerge. Methodological considerations for sampling personnel are discussed further in this section.

There is a general consensus that integrated systems should be tested with actual users. For example, Woods and Sarter (1993) state that because the final system should meet users' rather than the designer's needs, evaluation studies should involve users as test participants. They state that actual users may carry out tasks based on a thorough knowledge of and experience in the domain that others, such as domain-knowledgeable non-practitioners, may not have. ANSI (1992) states that design engineers should not be used as test participants, instead of actual users, because designers may have knowledge of the design or special skills in its use that tend to hide or underestimate design weaknesses. Meister (1986) stated, "It is a fundamental testing assumption that if the test is to be fully valid (i.e., predictive of operational system performance), it must be performed with personnel who are representative of those who will eventually operate and maintain the system" (p. 122).

Thus, participants should be representative of the personnel who will operate and maintain the plant. The logic to selecting personnel is essentially the same as that which is applied to selection of operational conditions. In general, the characteristics and demographic factors of the user population should be evaluated to identify those that can be expected to relate to task performance variability. Characteristics and factors that are expected to contribute to variability should be identified as dimensions to include in the sampling process. Several factors that should be considered in determining representativeness include:

- License and Qualifications - Separate selection criteria based on personnel qualifications should be established for supervisors, licensed operators, auxiliary operators, and other support personnel. The personnel roles should be consistent with staffing requirements for the actual plant and selection criteria should be representative of the characteristics of the intended user population of the plant.

- Skill/Experience - A range of skill and plant operating experience should be included to represent the experience levels of population potential users. For example, ANSI (1992) states that operational personnel selected to participate in operational tests should exhibit a mix of high and low skill levels to approximate the range of capability found with operational personnel. Baker and Marshall (1988) discourages the exclusive use of highly experienced and motivated participants because they "tend to perform well with almost any reasonable system" and, thus, can result in misleading and often artificially elevated levels of performance.

- Age - The distribution of user population age should be represented in the participant group.

- General demographics - Where applicable, characteristics such as physical size, motor and perceptual abilities, and cognitive capabilities (e.g., intelligence) should be representative of the range of these characteristics in the population of potential users.

- Shift Staffing - In selection of personnel, consideration should be given to crew composition. While the discussion of factors above is oriented toward characteristics of the individual participants, it is the crew that becomes the unit of analysis. Thus, selection of participants should include consideration of the assembly of operating crews, e.g., shift supervisors, reactor operators, shift technical advisors, etc., that will participate in the tests.

Once the important factors are determined, the population should be sampled to achieve a group of test participants that differ along the identified dimensions in a distribution similar to that of the population. A stratified random sampling process can be used to obtain an unbiased sample with the proper demographic profile.

To prevent bias in the sample, the following participant characteristics and practices should be avoided:

- Part of the design organization - These individuals can be expected to be bias towards the design and may have special knowledge about the design or the simulation that typical users would not have.

- Participants in prior evaluations - Participants should not be individuals that have participated in earlier tests and evaluations since they could be subject to the same bias that other contributors to the design process (Draft IEC, in preparation).

- Limited explicitly to volunteers - People who volunteer for studies may not be representative of the user population (ANSI, 1992).

- Selected for some specific characteristic, such as using crews that are identified as good or experienced.

A threat to statistical conclusion validity discussed in Section 3 was low sample size. To the extent that plant performance is dependent on the interaction of personnel with plant systems, human variability needs to be adequately represented in the data. Low sample sizes makes generalization difficult. As a general rule, the larger the sample size (number of participating crews), the greater the confidence in generalizing the observed test performance to predicted performance in the real world.

In general, the more variable personnel performance is, the larger the sample size that will be required to adequately represent human variability in integrated system performance. The actual sample size is difficult to specify precisely since it depends on several factors:

1. Covariation between personnel and system variability - The less sensitive the integrated system performance is to human performance, the less that variation needs to be assessed and the lower the needed sample size. For example, if an integrated system is automated to such a degree that

operator input has very little influence upon its performance, then it may not be necessary to include a large sample of personnel.

2. Crew homogeneity - To the extent that crew members are similar to each other along important personnel dimensions such as age and experience (as the result of selection criteria, for example), human variability will be reduced and the needed sample size will be lower.

3. Performance homogeneity - Factors such as high qualification criteria or rigorous training criteria will reduce human variability and lower the needed sample size.

4. Test design - The test design employed impacts the sample size needed. For example, a "between-subjects" design requires a larger sample size than a "within-subjects" design (see Section 5.7 for an explanation of test design).

It is unlikely that rigorous statistical approaches to sample size determination will be applicable to validation tests. [For scientific experiments, a power analysis (Cohen, 1969; Kramer and Thiemann, 1987) can be performed to identify the minimum number of participants necessary to achieve sufficient statistical power to reject the null hypothesis at a predetermined statistical error criterion.] Thus, the required sample size will be based on judgement and should consider the factors discussed above.

5.5 Operational Conditions

Section 11.4.4 of the HFE PRM states that integrated system validation should include dynamic evaluations for a range of operational conditions. In Section 4.2.2 of this report, System Representation Validity, the importance of representativeness of personnel was discussed. Operational conditions, like personnel, are an aspect of the integrated system that changes. Thus both fidelity and sampling need to be considered. The meaning of fidelity is similar to that discussed above for the HSI and process model. With regard to sampling, consideration should be given to characteristics of operational conditions that can reasonably be expected to cause variation in integrated system performance. The identified characteristics provide the dimensions that are important to performance. In research terms these dimensions would be referred to as independent variables. Those dimensions (independent variables) that are expected to contribute to system performance variation should be specifically identified and a sampling process should ensure that variation along that dimension is included in the validation. To the extent that the sampling process is appropriately conducted, representativeness of operational conditions is supported. When representativeness is compromised, threats to system representation validity emerge. Methodological considerations for sampling operational conditions are discussed further in the section that follows.

5.5.1 Operational Conditions Sampling

Operational conditions are described as combinations of plant states and configurations, events that will cause state changes, and situational factors. Operational conditions are developed in detail to generate test scenarios to be used as part of the validation test process. The operational conditions represented as test scenarios provide the context within which integrated system performance is evaluated. As discussed in Section 3, it is not possible to test every condition that is important to the actual operation of the plant. The simulated conditions encompass a finite and possibly small number of conditions in comparison to real-world conditions. Therefore, when selecting operational conditions,

a sampling process is necessary. The goal of sampling is to provide a basis to evaluate that the integrated system achieve its mission and to capture the dimensions that are likely to have important effects on integrated system performance (Chapanis and Van Cott, 1972) over the lifetime of the facility. The selection of operational conditions should provide a comprehensive basis to permit generalization to other conditions or combinations of conditions that were not explicitly addressed by the validation tests.

The operational conditions should concentrate on conditions that are important to safety, and should include conditions that are representative of the range of events that could be encountered during operation of the plant. The selected operational conditions should include, but not be limited to, the plant's design-basis conditions. The selection should go beyond design-basis conditions to support direct testing of feasible conditions that may not have been specifically addressed by designers (Vincente, 1992). The design basis for NPPs is single failure. But experience shows that multiple failures frequently occur. This is addressed through defense in depth of which operators are a key part. Thus, the test should ensure that the operator's capability to respond effectively to multiple failures (beyond design-basis events) and that performance of the integrated experiencing such an situation validated. For example, by testing a number of conditions at levels that have important effects on performance, inferences can be made regarding the adequacy of performance for scenarios that include levels of conditions and combinations of conditions that were not explicitly tested in the validation tests. However, if operational conditions are limited to design-basis events, then the degree to which test results can be generalized to other operational conditions will also be limited (Rasmussen, 1988; Woods and Sarter, 1993).

The sampling dimensions and scenario development considerations will be described below that should achieve an adequate representation of operational events with which to test the integrated system. In selecting the dimensions and scenarios, the test designer should consider the common sources of bias in the selection of test scenarios that were identified by ANSI (1992). This ANSI has cautioned that there is a general tendency for test designers to select scenarios with the following characteristics:

- Scenarios for which only positive outcomes can be expected.
- Scenarios that are relatively easy to conduct (e.g., scenarios that place high demands for simulation, data collection, or analysis are sometimes avoided).
- Scenarios that are interesting (e.g., which include HSI components or performance issues that are of particular interest).
- Scenarios that are familiar and well structured (e.g., which address familiar systems and failure modes that are highly compatible with plant procedures such as "textbook" design-basis accidents.

Sources of bias that reflect these characteristics should be addressed.

The following describes a several dimensions to guide the sampling of operational conditions. These dimensions are identified to meet the goals of operational condition selection identified above. These dimensions are essentially independent variables and reflect *characteristics* of operational events

and not individual test scenarios. One individual test scenario may reflect characteristics of many of the sampling dimensions. The sampling dimensions are grouped into three broad categories:

- Plant conditions
- Personnel tasks
- Situational factors that are known to challenge human performance

These sampling dimensions are not exhaustive nor are they entirely independent of each other.

Plant Conditions

The validation scenarios should include the following:

- Normal operational events including plant startup, plant shutdown or refueling, and significant changes in operating power.
- Failure events such as:
  - Instrument failures (e.g., safety-related system logic and control unit, fault tolerant controller, local "field unit" for multiplexer (MUX) system, MUX controller, and break in MUX line) including I&C failures that exceed the design basis, such as a common mode I&C failure during an accident.
  - HSI failures (e.g., loss of processing and/or display capabilities for alarms, displays, controls, and computer-based procedures).
- Transients and accidents such as:
  - Transients (e.g., turbine trip, loss of off-site power, station blackout, loss of all feedwater, loss of service water, loss of power to selected buses or CR power supplies, and safety and relief valve transients).
  - Accidents (e.g., main steam line break, positive reactivity addition, control rod insertion at power, anticipated transient without scram, and various-sized loss-of-coolant accidents).
  - Reactor shutdown and cooldown using the remote shutdown system.
- Reasonable, risk-significant, beyond-design-basis events.
  - These should be determined from the plant specific probabilistic risk assessment (PRA).

In selecting failures, consideration should be given to the role of the equipment in achieving plant safety functions (as described in the plant SAR) and the degree of interconnection with other plant systems. A system that is interconnected with other systems could cause the failure of other systems

because the initial failure could propagate over the connections. This consideration is especially important when assessing non-class 1E electrical systems.

Personnel Tasks

The scenario should reflect a range of interactions with HSI components and personnel:

- Range of risk-significant actions, systems, and accident sequences - The scenarios should test all risk-important human actions as defined by the task analyses and PRA and HRA, including those performed outside the control room. Situations where human monitoring of an automatic system is critical should be considered. Additional factors that contribute highly to risk, as defined by the PRA, should be sampled, including:

  - Dominant human actions (selected via sensitivity analyses),

  - Dominant accident sequences, and

  - Dominant systems (selected via important measures such as Risk Achievement Worth or Risk Reduction Worth).

- Range of procedure guided tasks - Regulatory Guide 1.33, Appendix A, contains several categories of "typical safety-related activities that should be covered by written procedures." The validation should evaluate selected activities based on procedures developed to address this guide. The evaluation should include appropriate procedures in each relevant category, that is,

  - Administrative procedures

  - General plant operating procedures

  - Procedures for startup, operation, and shutdown of safety-related systems

  - Procedures for abnormal, offnormal, and alarm conditions

  - Procedures for combating emergencies and other significant events

  - Procedures for control of radioactivity

  - Procedures for control of measuring and test equipment and for surveillance tests, procedures, and calibration

  - Procedures for performing maintenance

  - Chemistry and radiochemical control procedures

  Not all categories of procedures need to receive equal emphasis. Some categories (e.g., administrative procedures and procedures for performing maintenance) may be best evaluated as an adjunct to other tests. Administrative procedures are important to safe plant operation,

however, they may not need to be tested as completely as EOPs. Instead, selected situations governed by such procedures should be reflected in validation scenarios to ensure that such procedures, in conjunction with the rest of the integrated system, can achieve their intended functions without interfering with plant operations. Thus for example, situations involving equipment control (e.g., locking and tagging of equipment), shift and relief turnover, or maintenance of minimum shift complement and call-in of personnel, could be incorporated into selected test scenarios or validated separately.

Procedures for performing maintenance are least amenable to integrated system validation. While the design for maintenance is an important aspect of plant design, it does not typically involve validation of an integrated system. It is appropriate to validate maintenance that is to be performed in the control room while the plant is being operated. This validation should show that it can be accomplished without interfering with operator tasks that are necessary for monitoring and controlling the plant. Another aspect of maintenance to be validated is the capability of operators in the control room to control or track maintenance being performed in the plant.

- Range of human decision-making activities - The scenarios should reflect the range of activities performed by personnel, including:

  - Monitoring and detection (e.g., of critical safety-function threats),

  - Interpretation/diagnosis (e.g., interpretation of alarms and displays for diagnosis of faults in plant processes and automated control and safety systems),

  - Planning (e.g., evaluating alternatives for recovery from plant failures),

  - Execution (e.g., In-the-loop control of plant systems, assuming manual control from automatic control systems, and carrying out complicated control actions),

  - Obtaining feedback (e.g., of the success of actions taken).

  The range of scenarios should include tasks that exemplify skill, rule, and knowledge-based behavior (Rasmussen, 1986). Knowledge-based activities are particularly important and include activities for which personnel must use their knowledge of the plant to analyze contradictory evidence, test hypotheses, diagnose failures, plan courses of action, and evaluate consequences of planned actions. An example of a plant condition that may require knowledge-based behavior during diagnosis is a steam generator tube rupture with a failure of radiation sensors on the secondary side of the plant.

- Range of HSI components - The scenarios should address use of all types of HSI components:

  - Alarm system,

  - Display systems (e.g., discrete indicators, process displays, group-view displays),

  - Control systems: manual, automated, and combined manual and automated,

- Interface management facilities such as dialog design and navigation,
- Procedures,
- Job support and decision aids, and
- Communication equipment.

- Range of human interactions - The scenarios should reflect the range of interactions between plant personnel, including tasks that are performed independently by individual crew members and tasks that are performed by crew members acting as a team. These interactions between plant personnel should include:
  - Between main control room operators (e.g., operations, shift turnover walkdowns),
  - Main control room operators and auxiliary operators,
  - Main control room operators and support centers (e.g., TSC, EOF), and
  - Main control room operators with plant management, NRC, and other outside organizations.
- Tasks that are performed with high frequency.

Situational factors that are known to challenge human performance

The scenario should reflect a range of situational factors that are known to challenge human performance, such as:

- Difficult NPP Tasks - The scenarios should address the following categories of tasks that have been found to be problematic in the operation of NPPs, for example:
  - In-service surveillance testing and maintenance (e.g., equipment blocking, tagging, and bypass),
  - Procedure versus situation assessment conflicts,
  - Alarm management and secondary fault detection,
  - Fault detection, analysis, diagnosis, and mitigation,
  - In-the-loop control of plant systems such as feedwater control,
  - Detection of automated system failures and their override and manual control.

  The specific tasks selected should reflect the operating history of the type of plant being validated (or the plant's predecessor).

- Error-forcing contexts - Situations design to elicit human errors should be included in validation to assess the error tolerance of the system and the capability of operators to recover from errors should they occur. Most human errors can be explained on the basis of a relatively small number of cognitive mechanisms (Reason, 1988; Rasmussen, 1986). For example, Norman (1981, 1988) classified errors into three categories, based upon the cognitive mechanisms involved. Description errors result from the operator's characterization of a situation at too high a level of abstraction. This occurs because it takes less mental effort than constructing a detailed characterization. At such a high level of description, the operator may not have enough detail to select the appropriate actions. Premature diagnosis of a problem is an example of this type of error. The second type is activation or trigger errors, that occur when an intention leads to the activation of a schema, but the operator does not keep track of the resulting actions, or the automated sequence is interrupted in favor of another action. Failure to restore a valve to its proper position after maintenance may be an example of this type of error. The third type is capture errors that occur when the environmental cues are similar to those associated with a well-developed schema which is inappropriately activated. Changes in equipment or procedures in the CR make an operator susceptible to this type of error, if well-learned responses in the old CR are inappropriate in the new one. Also, similarity in the display of information patterns between two plant states can lead to capture errors. Another type of human error especially significant in digital systems is mode error (Sarter and Woods, 1995). Mode error is when the crew thinks a piece of equipment, component, or system is in one mode but it is really in another. This type of error has been associate with many accidents and near-misses in complex systems.

  The situational characteristics that increase the likelihood that cognitive error mechanisms will be utilized and, therefore, increase the likelihood of human errors may be called an error-forcing context (Barriere et al., in preparation). Error mechanisms and the forcing contexts for their occurrence have been described by numerous authors, including Norman (1988), Reason (1988), Rasmussen (1986), and Woods et al. (1994). The validation team should consider this literature in the development of error forcing context scenarios.

- High workload and multitasking - The coordination of concurrent tasks poses demands on individuals and crews for maintaining awareness of the status of individual tasks and shifting resources for task completion. An example may include concurrent performance of surveillance tasks during an operational event, such as a change in power.

- Workload transition - Human performance at any given time in a complex system may be effected by the level of workload for the preceding period of time (Huey and Wickens, 1993). Of particular interest are conditions that exhibit (1) a sudden increase in the number of signals that must be detected and processed following a period in which signals were infrequent and (2) a rapid reduction in signal detection and processing demands following a period of sustained high task demand (Smolensky and Hitchcock, 1993).

- Fatigue and circadian factors - Incidents involving human-machine interactions are more likely to happen when personnel become fatigued or when operating during the late night and early morning hours, such as on the backshift. Such factors contribute greatly to human performance errors. An effort to include scenarios that involve such factors should be included, such as inducing fatigue with long scenarios and conducting some tests on backshift hours.

- Environmental conditions such as poor lighting, extreme temperatures, and high noise that are known to degrade human performance (Echeverria, 1994).

5.5.2 Scenario Definition

The operational conditions selected for inclusion in the validation tests should be developed into detailed scenarios. Detailed scenarios represent combinations of the dimensions that were described above.

It is important that the scenarios have appropriate task fidelity so that realistic task performance will be observed in the tests and so that test results can be generalized to actual operation of the real plant. It is also important to have scenarios well defined so that they are replicated across participating crews with the exception of those aspects that change in response to crew behavior.

Baker and Marshall (1988) describe four factors that effect NPP simulator studies: test duration and operator experience, motivation, and expectancy. Simulation scenarios tend to be short in duration due to resource constraints (e.g., operator availability, simulator availability, test schedule and cost). As a result of the short test duration, simulated failures occur with high frequency during the course of the test session and participants have high expectancy for their occurrence. The short test durations interfere with the ability to simulate some performance shaping factors such as fatigue and boredom. This situation is quite different from real plant failure events, which can be characterized as long periods of passive monitoring interspersed with short intervals of intense activity. Reiersen, Baker, and Marshall (1988) stated that the test scenarios should have the following characteristics: (1) be long enough to deter the participants from constantly attending to the alarm status, and (2) the participants should not be aware that a transient will be inserted into the scenario and, thus, be "on guard" and predisposed to react in a particular way. Reiersen et al. addressed these considerations in their alarm study by designing test scenarios that were approximately 4 hours in duration, during which, the participants performed the task of running up a turbine from hot shutdown to 92% power. After the second hour of a scenario, transients were introduced to test the participant's use of the alarm system. The participants rated the realism of the test scenario "quite highly" and approval of even longer scenarios. This approach may be used in some validation scenarios to reduce participants' expectations of abnormal conditions. In addition, participants' expectations for the occurrence of abnormal conditions should be further reduced by the inclusion of scenarios addressing normal conditions in which no plant failures occur.

When developing test scenarios, the following information should be defined to ensure that appropriate performance dimensions are addressed and to allow scenarios to be accurately presented for repeated trials:

- Description of the scenario mission and any pertinent "prior history" necessary for operators to understand the state of the plant upon scenario start-up

- Specific start conditions (precise definition provided for plant functions, processes, systems, component conditions and performance parameters, e.g., similar to plant shift turnover)

- Events (e.g., failures) to occur and their initiating conditions, e.g., time, parameter values, or events

- Precise definition of workplace factors, such as environmental conditions
- Task support requirements (e.g., procedures and technical specifications)
- Staffing requirements
- Communication requirements with remote personnel (e.g., load dispatcher via telephone)
- Crew behavior requirements (e.g., information gathering, decision making, and plant control actions)
- Data to be collected and the precise specification of what, when and how data is to be obtained and stored (including videotaping requirements, questionnaire and rating scale administrations).
- Specific criteria for terminating the scenario.

When evaluating performance associated with the use of HSI components located outside the CR, the performance impacts of potentially harsh environments (i.e., high radiation) that require additional time should be realistically simulated (i.e., time to don protective clothing and access radiologically controlled areas).

The validation team should maintain an audit trail for each scenario which identifies the specific dimensions associated with each scenario. This will enable data analysts to trace problematic performance back to problematic scenarios and then back to the important performance dimensions they represent.

5.6 Performance Measurement

The concept of safety of integrated system performance was discussed in Section 4.2.3 Performance Representation Validity. Performance representation validity was described as the degree to which measures are representative of the performance characteristics that are important to safety. Because this concept is multidimensional, it is best represented by a comprehensive hierarchal set of performance measures.

Performance measurement is addressed in this section in terms of:

- The measurement characteristics that impact the quality of the performance measures,
- The identification and selection of variables to represent measures of performance, and
- The development of performance criteria.

5.6.1 Measurement Characteristics

This section describes 11 measurement characteristics that should be considered when selecting or developing performance measures. These characteristics are based upon several sources (ANSI/AIAA, 1992; Chapanis, 1972; Meister, 1986; Muckler and Stevens, 1992). Candidate measures

should be evaluated according to these characteristics. It should be noted that some may not apply to a particular measure of performance.

- *Construct Validity* - As was discussed in Section 4.2.3, Performance Representation Validity, a good measure is one that is representative of the performance domain that it is intended to represent. Many aspects of human performance are described in terms of hypothetical constructs, e.g., situation awareness and workload. Hypothetical constructs are not directly observed, they are inferred based on observation of behavior that is indicative of the construct. Thus, operational definitions must be developed for hypothetical constructs which describe the "operations" that must be employed to measure the construct. There are many possible operational definitions, some of which will be representative of the aspect of the performance of interest while others will not. For example, measuring situation awareness by counting the frequency with which one accesses an overview display is probably a poor operationalization of the construct. Asking the operator to identify plant status may be a better measure.

- *Reliability* - Refers to the repeatability of a measure and is a basic requirement of measurement. For example, if one measures the same behavior in exactly the same way under identical circumstances, the same measurement result should be obtained. To account for the intrinsic variability of human performance, the concept of reliability has been extended from the repeatability of a particular value to the repeatability of a measurement distribution. Thus, if one obtains the same measurement distribution with repeated measures, the metric is said to be reliable. Reliability is usually quantified using correlational statistics (Nunnally, 1967).

- *Resolution* - Measures should reflect the performance at an appropriate level of resolution, i.e., with sufficient detail to permit a meaningful analysis. For example, measuring operator movements in minute detail may not be appropriate if the evaluation concern is at a higher conceptual level, such as, "Was a particular plant system used as intended?".

- *Sensitivity* - Range refers to the score values that a measure can discriminate.Floor and ceiling effects that restrict variance should be avoided. Floor effects occur when the bottom of the scale range is not low enough to permit discrimination of lower scores. Ceiling effects occur when the top of the scale range is not high enough to permit discrimination of high scores. Frequency refers to the rate at which performance measures are taken. Performance measures should be sampled often enough to assess the behavior of interest. Includes the measure's range (scale) and the frequency of measurement.

- *Diagnosticity* - Refers to the characteristic of a measure that provides information that can be used to identify the cause of acceptable or unacceptable performance. For integrated system validation, measures of cognitive and task performance can add diagnostic value to the set of performance measures. They measure characteristics of human performance that may explain the observed plant performance.

- *Simplicity* - It is desirable to use simple measures both from the standpoint of executing the tests and from the standpoint of communicating and comprehending the meaning of the measures. However, simplicity should not be achieved at the expense of other considerations such as precision, reliability, validity, or generalizability.

- *Objectivity* - To the greatest extent possible, performance measures should be based on phenomena that are easily observed. This facilitates the assessment of the measure's reliability and the development of performance criteria. Muckler and Stevens (1992) have argued that the distinction between objective and subjective measures of human performance is not truly meaningful because all measures have some degree of subjectivity; i.e., susceptible to the personal biases of the individuals collecting the data and interpreting the results. While this is generally the case, objectivity, especially with regard to collecting data, is a desirable measurement characteristic.

- *Impartiality* - Measures should be equally capable of reflecting good as well as bad performance; i.e., issues or aspects of the HSI design that may reflect badly on the overall HSI design should not be avoided.

- *Unintrusiveness* - A measure may be considered unintrusive to the extent that its data-collection method does not significantly alter the psychological or physical processes that are being investigated. Webb et al. (1973) discussed the problems of the reactive effects of the measurement process itself on the data obtained; i.e., the degree to which the collection of data affects the behavior of the person or system that is being studied. Data collection methods that attract the participant's attention or disrupt the participant's activities may be problematic. For example, participants tend to behave differently when they know that they are being watched, particularly when the observer's actions provide the participant with feedback (e.g., certain actions by the participant result in the observer taking copious notes). Disruptions to participants' activities may result from (1) data-collection methods that restrict where and how a participant can act (e.g., participant must step around data collection equipment or stay within camera range) and (2) data-collection methods that impose additional demands on the participants (e.g., completing logs or questionnaires in the course of performing one's usual tasks). Less intrusive measurement methods may include naturalistic observation (in which the observer is invisible to the participants, as with a one-way viewing screen or hidden cameras) and automatic data collection (as in computerized logging of manual control actions). It may not always be possible, however, to completely avoid intrusiveness.

- *Acceptability* - Acceptance is a practical concern that is critical to obtaining valid measurement data. Poor acceptance of measures may result in participants refusing to cooperate, providing misleading data, or generally not taking the measurement process seriously. One consideration is confidentiality or protection from negative consequences associated with test performance. Participants may find some measures to be unacceptable if they feel that the measures will reflect negatively on their abilities. Another consideration is participant comfort. While the validation study may attempt to realistically simulate difficult work conditions, participant cooperation may be affected by factors such as stress, boredom, and physical discomfort associated with the test environment or the measurement tools (e.g., devices for physiological measures of stress). In addition, the manner in which measures are introduced to participants is important to their acceptance. Participants should be convinced that the measures address important issues, provide meaningful data, and that they will not experience negative consequences as a result of their participation in the test.

- *Administration* - Selection of performance measures includes considerations of the resources required to implement them, such as time, budget, personnel, equipment, logistics, and need for highly specialized expertise for data collection and/or analysis. The practical concern associated with performance can impact the quality of the data collected.

5.6.2 Variable Selection

In order to evaluate the performance of complex human-machine systems, it is necessary to adopt a comprehensive approach to performance measurement (see Figure 4.3). Accordingly measures of the performance of the plant (i.e., the HSI, components, systems, and functions) and personnel (i.e., primary and secondary operator tasks, cognitive factors, and physiological factors) should be obtained. In this section, the selection of performance measures will be considered.

Performance measures can include general and scenario-specific variables. General measures are those that can be applied to any scenario, independent of the details of the events of the scenario. Such measures might include critical safety function variables, workload scales, and physiological measures. While a specific variable might not be sensitive in a given scenario, it can never-the-less be meaningfully recorded.

Scenario-specific variables are dependent on the details of the scenario. Primary task measures may be a good example of these measures. A scenario may have a series of critical tasks that must be performed in order for the operational event to be successfully handled. Such tasks might be making a specific diagnosis, starting a specific pump, monitoring the actuation of an automatic safety system, or controlling steam generator water level within a tolerance band.

In the discussion below, emphasis is given to those variables that are general rather than scenario specific and to those reflecting personnel performance.

5.6.2.1 Plant Performance Measurement

As was illustrated in Figure 4.3, plant performance measures representing functions, systems, components, and HSI should be obtained. The higher the level of abstraction, the more general the measures. Thus, while measures of critical safety functions may be routinely obtained in all scenarios, measures of the status or performance of individual components will typically be highly dependent on the details of the scenario. Examples of plant performance measures that might be appropriate for a boiling water reactor loss of feedwater event are illustrated in Table 5.1. In this event (from NRC, 1987b), a loss of feedwater occurs by a trip of all feedwater pumps during power operations. Personnel actions, discussed in the next section are included as well. The crew must bring the plant to hot shutdown while maintaining water level above the top of active fuel elements. For each of the example performance measures listed in Table 5.1, criteria for which evaluating the measures should be established based on the specific characteristics of the scenario (see Section 5.6.3, Performance Criteria).

Table 5.1 Examples of Performance Measures for Loss of Feedwater

LEVEL	PERFORMANCE MEASURE
Function Level	
	Reactor core cooling/heat removal
	Rx temperature
	Rx water level
	Rx vessel pressure
System Level	
	HPCI
	flow rate
	RCIC
	flow rate
	RHR (SP cooling mode)
	SP temperature
Component Level	
	HPCI turbine
	lube oil temperature, pressure
	RCIC turbine
	lube oil temperature, pressure
	RHR valves (for SP cooling mode)
	position
HSI Level	
	Alarms
	Rx water level
Personnel Actions	
	Time to execute "Monitor and
	Control RPV water level"
	procedure (errors)
	Time to verify autostart or start of

HPCI (errors)

Time to verify autostart or start of

RCIC (errors)

Time to restore RPV water level to

normal band using HPCI/RCIC

(errors)

Time to initiate SP

cooling (errors)

Time to secure HPCI when RPV

water level reaches normal band

(errors)

RCIC = reactor core isolation cooling system; RCS = reactor coolant system; RHR = residual heat removal;
RPM = revolutions per minute; RPV = reactor pressure vessel; Rx = reactor;
SP = suppression pool

5.6.2.2 Personnel Task Measurement

Measures of personnel task performance provide data that compliment the plant performance measures. For example, even when plant performance measures are maintained within acceptable ranges, shortcomings in the design may result in unnecessary demands being placed on operators. These demands will be manifested in the operators' behavior, e.g., the accuracy and timeliness of event detection and decision making. Measures of task performance can reveal potential human performance problems that were not reflected in plant performance measures in the test and evaluation context.

Personnel task measurement can be conceptualized along several dimensions: task type and method of identification. Two types of tasks are significant: primary tasks and secondary tasks. Primary tasks are those involved in performing the functional role of the operator to supervise the plant; i.e., process monitoring, decision-making, and control. Secondary tasks are those the operator must perform when interfacing with the plant, but which are not directed to the primary task.

The second dimension is method of identification. Quantification of personnel tasks should be performed using both top-down and bottom-up approaches. For a top-down perspective, the tasks that personnel should perform must be identified for each specific scenario (see Table 5.1). Such tasks can include necessary primary (e.g., start a pump) as well as secondary (e.g., access the pump status display) tasks. Top-down analysis also facilitates the identification of errors of omission by identifying tasks which should be performed. From a bottom-up perspective, the tasks that are actually performed by personnel during simulated scenarios should be identified and quantified. That is, while the top-down perspective will identify tasks that should be performed, it is important to identify tasks that are actually performed. While these should include the required tasks, they will include others as well. It may be possible to anticipate the complete set of required actions in advance of conducting the tests. However, the set of actual tasks may be somewhat different from those anticipated in top-down analysis because of the interactions between personnel actions (or inactions) and plant dynamics. The bottom-up approach will enable a post hoc analysis of these effects of interactions. It will also facilitate the identification of errors of commission.

With respect to primary tasks, procedure steps may serve as a guide to identifying a set of tasks to measure. To supplement procedures or provide task information in the absence of procedures, task analysis may be used. Consideration should be given to the level of detail that should be obtained. For example, for some simple scenarios, measuring the time to complete a task may be sufficient. For more complicated tasks, especially those that may be described as knowledge-based, it may be appropriate to perform a more fine-grained analysis such as identifying task components: seeking specific data, making decisions, taking actions, and obtaining feedback. Tasks that are critical to successful integrated system performance and are knowledge-based should be measured in a more fine-grained approach.

With respect to secondary tasks, the specific measures of the demands imposed by the design will depend on the detailed implementation. However, activities such as the following may be obtained:

- Configuring the workstation, e.g., adjusting monitors and keyboards
- Selecting mode configurations for computer support functions or equipment

- Navigating between displays
- Navigating within displays
- Formatting and manipulating displays (e.g., changing display type and setting scale)
- Searching for procedures
- Searching through procedures
- Searching for controls
- Performing activities such as ad hoc job support aid usage (e.g., placing markers in procedure pages or placing tape on an indicator to mark information)

Meister (1985) has developed a general taxonomy that may be used to quantify primary and secondary task performance. The taxonomy includes variables such as:

- Time
  - Reaction time, e.g., time to perceive event, initiate action, initiate correction, detect trend of multiple-related events
  - Time to complete activity
  - Overall task time (duration)
  - Time sharing among events
- Accuracy
  - Correctness of observation, i.e., identifying stimuli internal and external to system detection of changes or trends, recognition of signal in noise, recognition of out-of-tolerance condition
  - Response correctness, i.e., accuracy in control positioning, display reading, decision making, communicating
  - Error characteristics, e.g., amplitude and frequency measures, content analysis, change over time
- Frequency
  - Number of responses per unit, activity, or interval, e.g., control and manipulation responses, communications, personnel interactions, diagnostic checks
  - Number of performance consequences per activity, unit, or interval, e.g., number of errors, number of out-of-tolerance conditions
  - Number of observing or data-gathering responses: observations, verbal or written reports, requests for information, rate of engagement
- Amount achieved or accomplished
  - Degree of success
  - Percentage of activities accomplished
  - Measures of achieved performance (e.g., terminal or steady-state value)

- Consumption or quantity used
  - Total resources consumed
  - Resources consumed as a function of time

- Subjective reports of participants

- Behavior categorization by observers
  - Judgment of performance: rating of operator and crew performance adequacy, rating of task or mission segment performance adequacy, estimation of amount (degree) of behavior displayed, measures of achieved maintainability, equipment failure rate (mean time between failures),cumulative response output,proficiency test scores (written)
  - Magnitude achieved: terminal or steady-state value (e.g., temperature high point), changing value or rate (e.g., degree of changes per hour)

Particular performance measures should be chosen to reflect the important aspects of the task with respect to system performance. For example, the time taken to respond to a given indication will not provide meaningful information if it is applied to a situation in which neither the state of the plant nor the procedures to be followed mandates an immediate overt response. When task analyses indicate that coordination or communication among operators is required to complete a task, measures of task performance that are defined in terms of the crew (rather than an individual) should be provided. In addition, global measure should be considered, e.g., HSI "overhead" (time spent engaged in secondary tasks as a function of total time available).

In addition to describing task performance, the assessment of performance should also focus on capturing human errors in both primary and secondary tasks. Again, top-down and bottom-up logic applies. When specific performance criteria can be defined, identifying errors is relatively easy. For example, if the performance criterion is to initiate standby liquid control within 10 minutes, then failure to do so can be defined as an error. As a second example, if a specific sequence of procedural steps must be followed, deviations from the sequence may be an error. However, not all errors can be easily identified in advance. Thus, task performance should be carefully observed so that other errors, which could not be predefined, may be detected.

5.6.2.3 Cognitive Factors Measurement

Supervisory control consists, in large part, of cognitive processes, e.g., monitoring, decision making, and control. In fact, most operator errors can be explained on the basis of a relatively small number of cognitive mechanisms that reflect the operator's response to high information and high complexity situations that require controlled information processing and place high demands on attentional resources and working memory (Norman 1981, 1988; Reason 1988, 1990). Factors such as situation awareness and cognitive workload underlie the operator's primary task performance. The measurement of situation awareness and cognitive workload will be described in the sections to follow.

5.6.2.3.1 Situation Awareness

Many different definitions of situation awareness have been discussed in the literature and many relate situation awareness and mental models. In the present discussion a distinction will be drawn between the two concepts (see O'Hara, 1994 for a more in-depth discussion of these concepts).

The knowledge governing the performance of highly experienced individuals may be referred to as a mental model and constitutes the operator's internal representation of the physical and functional characteristics of the system and its operation. It is built up through formal education, system-specific training, and operational experience. The mental model is represented in knowledge or long-term memory. An accurate mental model is considered the defining characteristic of expert performance in general (e.g., Wickens, 1984) and for NPP operations in particular (e.g., Bainbridge, 1986; Moray, et al., 1986; Rasmussen, 1983; Sheridan, 1976). The mental model is thought to directly drive skill-based processing, to control rule-based activity through the mediation of the operator's conscious effort in working memory, and to provide the substantive capability to reason and predict future plant states required of knowledge-based processing (Rasmussen, 1983). Moray (1986) argued that a well-developed mental model enables the operator's performance to become more "open-loop" and thus, system control to become smoother. "Open-loop" in this context means that behavior becomes less driven by feedback and more governed by the operators prediction of future system behavior and the desired goal state. The mental model allows prediction and expectancy to guide control responses; however, expectancy can also make the detection of subtle system failures difficult (Wickens and Kessel, 1981). Similarly, Bainbridge (1974) stated that the operator of a NPP uses the mental model to predict the near-term future state of the plant and then uses this inference to guide sampling of indicators to confirm the inference.

By contrast to the relative permanent characteristics of the mental model, an operator's current interpretation of a system's status may be referred to as situation awareness. Situation awareness is the degree of correlation between the operator's understanding of the plant's condition and its actual condition at any given time. An operator can have a good mental model (e.g., knowledge of how the plant functions) but poor situation awareness (understanding of its current status). Situation awareness has also been identified as the single most important factor in improving crew effectiveness in complex systems (Endsley, 1988).

Theoretical treatments of situation awareness (Endsley, 1993b, 1995a; Fraker, 1988) suggest that it typically depends heavily on working memory. In addition to supporting situation assessment and projection of future status, working memory must also support other functions, e.g., the selection and execution of operator actions. Accordingly, if a task places high demands on working memory, situation awareness may suffer. On the other hand, the demands placed on working memory in maintaining situation awareness can be lessened if the situation is familiar (based on training or experience) and can, therefore, be identified with a representation stored in long-term memory. In this case, it would not be necessary for the operator to maintain in working memory each detail of the situation.

Based on this brief discussion, it is clear how situation awareness and cognitive workload (see next section) may vary inversely under complex, somewhat ambiguous situations. For example, under unfamiliar or otherwise difficult conditions, high cognitive workload may be associated with decreased situation awareness owing to lack of available working memory resources. However, as Endsley (1993b) points out, situation awareness and cognitive workload, while inter-related, may vary independently such as when a task is intensive, but readily recognizable. This is because situation awareness requires the expenditure of cognitive resources that contribute to workload, but it is not the only cognitive activity requiring such resources. Thus workload and situation awareness are separate concepts.

Situation awareness can be assessed by three general methods:

- Performance-based techniques - Situation awareness is inferred based on the performance of the system and/or operator,
- Subjective rating techniques - Situation awareness is rated by the operators themselves,
- Direct query techniques - Situation awareness is revealed by questioning operators about their knowledge of particular aspects of the situation.

Each of these is discussed below. (See also Adams, Tenney, and Pew, 1995, for a discussion of the strengths and weaknesses of different approaches to situation awareness measurement.)

Performance-Based Techniques

Poor system performance is often a result of inaccurate assessment of situations by operators. However, it is not necessarily the case that poor system performance indicates poor situation awareness (Endsley, 1993a). Further system performance may remain within acceptable limits despite poor situation awareness. The utility of assessments of situation awareness based on operator performance (as interpreted, e.g., by expert observers) is similarly limited, since not all aspects of the operator's knowledge are available to the observer (i.e., reflected in behavior or verbalizations).

Techniques intended to reveal to observers the content of the operator's awareness should be used with caution because their effectiveness may be limited and, more importantly, because they may alter the operator's ongoing task. For example, information on a particular display can be removed or altered in order to assess an operator's awareness of that information. However, an operator may not overtly react to the alteration immediately (or at all), e.g., if the operator assumes that an instrument has failed.

Alternatively, techniques might be employed to prompt verbalization by the operator regarding aspects of the situation that are of interest. Sarter and Woods (1991) suggest that the intrusiveness of this technique can be minimized if the prompts are task relevant, e.g., situations may be defined in which it is necessary for the operator to inform others, such as, fellow operators or technical support personnel, of the status of particular systems. However, such techniques may draw the operators' attention to aspects of the situation to which they otherwise would not have attended.

In conclusion, performance-based techniques have both logical ambiguities in their interpretation and practical problems in their administration. Thus, they may not be well suited as measures of situation awareness in validation tests.

Subjective Ratings Techniques

Having operators subjectively rate their situation awareness is comparatively more simple and direct than performance-based techniques because it does not rely as heavily on inference. The use of operator ratings can also be less intrusive. Because ratings are typically made after completion of an exercise, there is no need to alter or interrupt the operator's ongoing tasks.

A multi-dimensional subjective Situation Awareness Rating Technique (SART) has been described by Taylor (1989). In the "3-D" version of SART, scenarios are rated by operators on three dimensions:

- Demand on attentional resources (instability, complexity, variability of the situation),
- Supply of attentional resources (arousal, concentration, division of attention, spare mental capacity), and
- Understanding of the situation (information quality, information quantity, familiarity).

If greater diagnosticity is desired, ratings may be elicited on each of the ten component constructs (in parentheses above); this is referred to as the "10-D" version.

Vidulich and Hughes (1991) describe an adaptation of the subjective workload dominance (SWORD) metric (Vidulich, 1989) for the assessment of situation awareness; the technique is referred to as situation awareness (SA)-SWORD. Like the SWORD technique, SA-SWORD is based on retrospective relative evaluations of all conditions that are structured and analyzed according to a variant of the Analytic Hierarchy Process (Saaty, 1980). In a test of SA-SWORD, Vidulich used the same data collection and analysis procedures as for SWORD, changing only the instructions. The authors note that half of the subjects apparently equated situation awareness with the information demands of the task rather than indicating the degree to which the demands could be met. They conclude that the sensitivity of SA-SWORD has not been conclusively demonstrated and that it will be necessary to more clearly define the concept of situation awareness to subjects who are to provide the subjective ratings.

One problem with these methods is that the they include workload factors rather than limiting the techniques to situation awareness measurement itself; i.e., as an independent construct. For example, two of the three constructs which make up the SART metric have traditionally been included in definitions of workload (demands on attention and supply of attentional resources). The third, understanding of the situation, is more closely related to the concept of situation awareness, but it could be argued that only the familiarity subscale relates uniquely to situation awareness. Techniques based on post hoc ratings also have the disadvantage that they necessarily depend on operators' recall of their situation awareness, which may be bias and by the outcome of the exercise. Furthermore, the operators rate their awareness of the situation as they perceive (or reconstruct) it, not relative to objective information (Endsley, 1993a).

In conclusion, subjective-rating techniques also have logical ambiguities in their association with workload and accuracy limitations associated with the need to recall awareness information. Thus, they may not be well suited as measures of situation awareness in validation tests.

Direct Query Techniques

Direct query techniques involve questioning operators about their knowledge of particular aspects of the situation. When the questions are asked after a scenario, it is naturally not intrusive, but it suffers from the same limitations as post-exercise ratings, i.e., it depends heavily on memory and may be influenced by the outcome of an exercise. Questioning the operator about aspects of the

situation while an exercise is in progress can avoid some of the above difficulties, but it is necessarily intrusive. In this case, responding to questions is in effect a secondary task for the operator which can disrupt performance (Endsley, 1993a).

The Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988, 1995b) has been developed to avoid the above limitations. The technique takes advantage of the fact that a simulator exercise can be stopped at any point and then continued. At a random point in time the simulation is stopped and all displays go blank. Operators can then be asked a series of questions related to their current awareness of the situation. When the simulation is completed, the operator's responses can be compared with what was actually happening at the time the simulation was halted. Dynamic changes in situation awareness can be detected by including several data collection stops during a single scenario. The content of the questions and the times at which they are to be asked must not be predictable by the operator to avoid altering the task (e.g., changing the operator's sampling of instrument readings).

Although it has been argued (Sarter and Woods, 1991) that the SAGAT technique is highly intrusive, studies performed to assess intrusiveness fail to demonstrate disruption of task performance (Endsley, 1993a, 1995b). The methodology has also been used successfully in experiments with nuclear plant operators (Hogg, Folleso, Torralba, and Volder, 1994).

Another potential problem is that the questions posed in situation awareness sessions may cue operators to details of the scenario. This can be avoided by imbedding key situation awareness questions in with other, less relevant questions.

Finally, SAGAT assessments may be influenced by memory effects, i.e., operators must recall information to answer questions and, therefore, weakness or distortions which result from recall processes may make the assessment of situation awareness less accurate. However, such effects should be less corrupting than for other available techniques.

In conclusion, direct query techniques have been used successfully in numerous applications including simulated nuclear plant operations. When used, the application should be carefully designed to minimize the intrusiveness of the question sessions by keeping them short and infrequent. Also, the potential for cuing operators by question presentation should be minimized. The methodological considerations for use of this technique are discussed by Hogy, et al. (1994) and Endsley (1995b).

5.6.2.3.2 Cognitive Workload

Cognitive workload has important relationships to human performance and error. Despite its importance and intuitive appeal, a precise, generally accepted definition of cognitive workload is still not available (Wickens, 1992; Moray, 1979; Hancock and Meshkati, 1988). This is mainly because workload is a multivariate concept, reflecting the effects of environmental and situational factors on information processing as well as the operator's perception of subjective effort and stress. As such, workload is best thought of as a hypothetical construct having many possible operational definitions depending on the context in which it is being evaluated. Thus, many different methods have been proposed to assess cognitive workload. A through review of all the methods is beyond the scope of this report, but the reader is referred to Wierwille and Williges (1978), Moray (1979), O'Donnell and

Eggemeier (1986), Hancock and Meshkati (1988), Lysaght et al. (1989), and Wierwille and Eggemeier (1993) for reviews of the various alternatives.

Techniques for measuring cognitive workload can be divided into two broad types: predictive and empirical (Lysaght et al., 1989). Both types of techniques are discussed below.

Predictive Techniques

Predictive techniques use either analytic or projective methods (Vidulich, Ward, and Schueren, 1991). Analytic techniques include mathematical modelling, task analysis, and simulation modelling; all of these methods require detailed specifications of the operation of the system. Projective techniques do not require a detailed description of the system to be employed. These techniques depend on the judgements of subject matter experts (SMEs) and comparisons with existing systems.

There are types of methods grouped into the class of analytic techniques (Lysaght et al., 1989). Most of the efforts to predict operator workload based on mathematical models have applied models from one of three theoretical domains: manual control theory, information theory, and queuing theory. Manual control models are best suited to tasks involving continuous manual control and are, therefore, unlikely to be of use in the assessment of workload in supervisory control settings. Techniques based on information theory and queuing theory can be used to identify periods of high information transfer during a tasks, but their correlation with workload may not always be high (Lysaght et al., 1989).

Estimates of cognitive workload can be made based on task analyses by analyzing the time required to perform tasks in comparison with the time available. This approach can be used to identify gross deficiencies in a design. When more precise workload assessments are required, tasks can be partitioned into smaller components (e.g., sensory channels or cognitive capacities).

Computer simulation techniques are increasingly being applied to the analysis of human operators in complex systems. Task analyses serve as a basis for the construction of task network models of operator activity; these models are then implemented in computer simulation languages which have the capability to represent operator behavior. See Meister (1985) and Chubb, Laughery, and Pritsker (1987) for detailed treatments on the methods involved in task network modeling.

Projective techniques utilize opinion. Given a general system description, operator workload may be predicted by elicitation of expert opinion and comparison with existing systems (Lysaght, et al., 1989). Expert opinion allows the identification of broadly defined workload problems. However, the results are very subjective and differences among SMEs may be large. Techniques have been developed to help structure expert opinion in order to make results more reliable and useful. The use of projective workload rating techniques (see the discussion of subjective workload measures later in this section) is an example of such an approach (Reid, Shinledecker and Eggemeier, 1981). The quality of workload predictions can be further enhanced to the extent that SMEs can base their judgements on experience with existing systems.

SMEs can also be provided with detailed task analysis information for the new system (as described above) and use this information to make estimates of the component workloads (e.g., for visual, auditory, and decision workload). The combination of detailed task analysis information and

SME judgements has been utilized in the development of military aircraft. The approach becomes less applicable the more a new design deviates from the one that operators are familiar.

Empirical Techniques

Empirical workload assessment approaches are typically divided into four classes of measurement technique (Wierwille and Williges, 1978): spare mental capacity measures, subjective measures, physiological measures, and system performance measures. The use of system performance measures to evaluate workload is not discussed in this document because (1) they are not always suitable for use with highly trained operators (as was discussed in Section 3.2.3), (2) more direct measures of workload are available, and (3) workload is logically treated as a personnel measure that supports system performance.

Spare Mental Capacity Techniques

The primary assumption of this method is that the human operator has a finite quantity of resources available to process information. That portion of the capacity not required for performance of a primary task is available as spare capacity for other subsidiary tasks. The workload imposed by a primary task may be determined by measuring the speed and accuracy of a subsidiary task since it can only be performed with the mental capacity not used by the primary task. The more capacity needed for the primary task, the less is available for the subsidiary task and it will be performed less well. When performance on the primary task is maintained, a decrease in performance of the subsidiary task reflects higher workload of the primary task.

A multiplicity of tasks have been used to assess spare capacity (see ANSI, 1993 and Eggemeier and Wilson, 1991 for a brief description of numerous subsidiary tasks). The most frequently used tasks are 1) choice reaction time, 2) time interval estimation and production, 3) memory search, and 4) mental mathematics.

One of the assumptions implicit in the use of spare capacity measures is that resource expenditure associated with the primary task is not changed by the introduction of the subsidiary task (e.g., the operator does not devote fewer resources to the primary task after the subsidiary task is introduced). Therefore, intrusiveness is a key consideration in choosing a subsidiary task for workload assessment in the test and evaluation context. If subsidiary tasks are used, they should be selected from among those that have been shown not to degrade primary task performance (see Lysaght, et al., 1989, also reproduced in ANSI, 1993, for a summary of the interactions that have been reported in the literature).

Another assumption is that the primary and subsidiary tasks draw on the same pool of resources. This implies either that 1) there is a single, undifferentiated resource, or 2) the subsidiary task is selected to tap the resources that support the primary task. Current theoretical treatments of workload favor the second implication in that they assume multiple resources (Wickens, 1980). Subsidiary tasks should be selected to tap into the cognitive resources of interest, e.g., perceptual versus control processes. In the context of supervisory control, a subsidiary task that taps into the processing resources associated with working memory is highly desirable due to its important role in developing situation awareness, decision-making, and human error (Fraker, 1988; O'Hara, 1994).

Knowles (1963) identified additional subsidiary task properties that are important considerations. First, operators should be able to learn to task easily; there should not be wide variations among operators in their ability to perform the task. Second, the task should be self-paced;

that is, the operator should be able to adjust the amount of attention devoted to the task as required by the demands of the primary task. Third, the scores that describe performance should be comparable across situations, and it should be possible to score performance continuously during an exercise.

It is sometimes possible to use a subsidiary task to be similar to tasks that an operator might perform (in addition to the primary task) during actual operations. Operator acceptance and compliance with subsidiary procedures will be greater if the tasks are introduced in operationally relevant ways. For example, operators may be asked to obtain certain parameter readings in support of a maintenance task. The instructions can emphasize that the information should be obtained only when there is time to do so. The length of time to respond to such requests may serve as a measure of how demanding the primary task is. While tasks presented in this way may be less artificial, they may nevertheless be intrusive since they are not in fact part of the operator's actual task.

Subjective Techniques

Subjective workload measurement techniques utilize operator ratings to assess workload. The specific aspects of cognitive workload that are assessed depend on what the operator is asked. Some subjective measures provide assessments of workload along global dimensions. The best known global subjective measure is the Modified Cooper-Harper (Wierwille and Casali, 1983) scale. The operator rates the difficulty level of the task and the level of mental effort required on a 10-point scale.

Subjective rating scales have been developed to assess specific components of workload as well. This type of rating scale offers the potential for diagnosticity in measuring workload. While the theoretical foundation of the use of subsidiary task approaches to workload assessment are relatively clear (spare mental capacity measurement), the specific sources of subjective measurements of workload as less straightforward. Several investigations have attempted to identify common factors influencing operators' subjective workload ratings and to develop measures of them.

Jex (1988) compared several different approaches to subjective workload ratings and identified three dominant factors: busyness (attention demands), task complexity (the cognitive difficulty of task components), and consequences (concern or importance of one's task performance to mission success). Reid and Nygren (1988) performed a more comprehensive comparison of subjective workload techniques and identified three dominant factors which were quite similar to those proposed by Jex. They were time load (e.g., time stress and time required versus time available); mental effort load (e.g., perceived effort and task complexity), and psychological stress load (e.g., fatigue and uncertainty and risk). These three subjective workload dimensions were developed into the Subjective Workload Assessment Technique (SWAT). SWAT employs three-point rating scales for the three dimensions.

The National Aeronautics and Space Administration (NASA) has conducted an extensive program to identify the major factors that contribute to subjective workload assessments (Hart and Staveland, 1988). Their studies were based upon empirical analysis of the rating scales of various tasks. Analyses of rating intercorrelation matrices revealed six principal factors: mental demand, physical demand, temporal demand, own performance, effort, and frustration. These dimensions were developed into one of the most frequently used workload assessment techniques - the NASA Task Load Index (TLX) (Hart and Staveland, 1988). Operators rate the six workload dimensions on 20-point rating scales.

The factors identified by SWAT and the NASA TLX are very similar providing credibility to their identification of important factors in subjective workload evaluations.

Each of the techniques identified above, are based upon operator's use of rating scales to assess an integrated system. A less widely used subjective workload measurement technique is based on relative judgement. Rather than rating a single system, operator retrospectively judge a system relative to another one with respect to the workload experienced. Workload ratings are generated from the judgement matrix which is produced by means of computational algorithms, e.g., the Analytic Hierarchy Process (Saaty, 1980). The Subjective Workload Dominance (SWORD) technique (Vidulich, 1989) is an example of the relative judgement approach.

Subjective measures are typically unobtrusive and often very sensitive (Wierwille and Williges; 1978, Williges and Wierwille, 1979). The sensitivity of subjective methods has been demonstrated in a number of studies. Global workload assessment techniques have demonstrated sensitivity to workload in flight simulation and remotely-piloted vehicle control environments (Eggemeier and Wilson, 1991). The NASA-TLX and SWAT have demonstrated sensitivity in a variety of multi-task environments (Eggemeier and Wilson, 1991). SWAT has also demonstrated sensitivity in a simulated NPP setting (Beare and Dorris, 1984).

Hicks and Wierwille (1979), for example, compared five methods of workload measurement in the assessment of the difficulty of simulated automobile driving tasks. Tasks were varied in difficulty by changing placement of gusts of wind. Only two of the five methods successfully discriminated task difficulty: subjective ratings and primary task importance.

Vidulich (1988) concluded that subjective measures are especially sensitive to information processing load and for "evaluating the impact of automation on operators serving primarily as system monitors," where the greatest demand is on the operator's decision-making capabilities. Subjective measures are also considered the most acceptable to operators of all workload measurement techniques (Wickens, 1984).

Several conclusions emerge from this research. The advantages of using subjective measures include:

- Subjective workload ratings may provide a more general and comprehensive assessment than spare capacity methods. Thus, they may provide an excellent complement to spare capacity task methods and can be used in a variety of test settings.

- A fairly consistent set of factors have been identified across several different investigations as contributing to subjective workload ratings.

- Subjective measures have been found to be very sensitive to workload and especially appropriate to tasks involving monitoring and supervisory control of mainly automated systems.

- Subjective measures are typically unobtrusive and acceptable to operators.

Several unanswered questions and disadvantages of using subjective measures have also been noted in the literature (Williges and Wierwille, 1979):

- It is often difficult to distinguish mental from physical workload. This problem is minimized to the extent that the task under investigation has minimal physical work involved.

- Subjective ratings of workload can be influenced by "emotional state, experience, learning, and natural abilities." A strong psychometric instrument would help minimize these influences.

- Adaptivity of operators to the task can alter ratings. For example, uniqueness of a simulator may make a task initially seem more difficult and when the operator adapts to the simulator, the task may seem easier than it ordinarily would.

- If subjective ratings are obtained following a scenario, they may be subject to memory effects; i.e., the workload associated with early and late phases of a scenario may exert more influence on ratings than the middle of the scenario (O'Hara, 1994).

- Subjective ratings require conscious knowledge of how demanding a task was. The degree to which all workload-related factors are consciously experienced by the operator (and therefore available to be reported) is unknown.

Physiological Workload Assessment Techniques

Several physiological processes have been investigated as potential indicators of task demand. Physiological measures of workload involve the measurement of physiological parameters, such as heart rate, which are thought to covary with workload levels. Since multiple measures of each process have been considered, the are numerous techniques described in the literature. A general review of this literature is presented in O'Donnell and Eggemeier (1986). Kramer (1991) and Wilson and Eggemeier (1991) review the literature from the perspective of multi-task performance. The most frequently employed categories of physiological measures of workload are cardiac measures, brain activity, and eye activity (Wierwille and Eggemeier, 1993). Each is briefly discussed below.

Cardiac Measures. Heart rate has been shown, in both laboratory and operational environments, to increase with the overall arousal, physical exertion, and/or emotional responses associated with task demands (Wierwille and Eggemeier, 1993). However, a number of studies failed to demonstrate a systematic relationship between workload and heart rate, leading to speculation that different types of task demands may have opposite influences on heart rate (Kramer, 1991).

Heart rate variability has also been examined as a indicator of workload; variation in beat-to-beat intervals has been shown to decrease with increasing workload. Measures based on spectral analyses of heart rate variability (particularly the power in the 0.10 Hz component) have also been investigated and found to be systematically related to workload. However, the relationship is generally found for relatively large differences in task demands (Kramer, 1991) and its usefulness in multi-task environments has not been demonstrated (Wilson and Eggemeier, 1991).

Like many physiological measures, cardiac measures require the attachment of sensors to the participants which may cause discomfort and lack of operator acceptance. Collection of cardiac measures typically involves the use of electrodes. Fortunately, the placement of the electrodes is not critical because of the large signal-to-noise ratio of the electrocardiogram (ECG). Portable devices for recording the ECG are available. Devices that measure blood volume in tissue (e.g., photoelectric

sensors worn on the ear or finger) can also be used for purposes of heart rate recording (Kramer, 1991). However, cardiac measures are unintrusive in that they do not require the introduction of additional stimuli or response requirements.

Measures of Brain Electrical Activity. Two classes of brain activity measures have been used as workload indicators: ongoing electroencephalographic (EEG) activity and evoked potentials (EP). Ongoing EEG is analyzed by determining the Fourier components of the electrical activity and calculating the power at each frequency. The power in different frequency bands, particularly the alpha (8-13 Hz) and theta (4-7 Hz) bands, has been shown to be sensitive to workload differences. There is no evidence that EEG activity is selectively sensitive to specific processing demands; rather it reflects overall arousal or alertness (Kramer, 1991). Like heart rate, EEG will likely reflect emotional and physical load as well as information processing demands.

While technological developments have made it possible to collect EEG without tethering test participants to amplifiers, analysis of the signals still requires equipment and expertise. Furthermore, in operational testing environments the EEG may be subject to contamination by electrical noise from equipment and by electrophysiological noise generated by the operators themselves (especially if they are moving about or speaking to other operators). Electrical filtering of signals is necessary to minimize the effects of such noise.

The EP is electrical activity associated with a specific event. It is obtained by averaging a number of samples of EEG that are time-locked to that event, thus "averaging out" the ongoing EEG. The response consists of a number of positive and negative peaks occurring within 750 msec of the presentation of a stimulus (Wilson and Eggemeier, 1991). Of particular interest is a positive wave that occurs at roughly 300 msec (P300) in response to rare task-relevant events. It has been argued that the P300 wave is associated with the updating of a mental model (Gopher and Donchin, 1986). Insofar as the P300 is not evoked by unattended stimuli, it is a potential indicator of workload. The amplitude of the P300 recorded while the primary task is being performed will reflect the amount of attention that remains available. The lower the amplitude of the P300 wave, the greater the attention demanded by the primary task.

There are a number of potential disadvantages to using EP techniques in operational testing. Because the EP is a very small signal it is subject to electrical artifacts. The greater the noise, the larger the number of samples that are required to produce a useful averaged signal. In addition, the EP typically requires the repeated presentation of an evoking stimulus and, usually, a covert or overt response by the operator. Thus the EP technique can be intrusive in the same manner as subsidiary task techniques. Sampling can be time-locked to operationally relevant events, but it may not be plausible to present such events as often as is necessary to produce a useful averaged EP (Wilson and Eggemeier, 1991).

Measures of Eye Activity. Three aspects of endogenous eye blinks (blinks that are not reflexive responses to environmental stimuli) have been evaluated as measures of workload: blink rate, blink duration, and blink latency. Measures can be collected by means of electro-oculogram (EOG) recording or by analysis of videotaped activity. Both blink rate and latency typically decrease as the amount of visual information to be processed increases. This relationship is not consistently found, however. Blink duration is also observed to decrease with increasing visual demand. Blinks increase in duration with time on task, presumably due to fatigue (Kramer, 1991).

It has been suggested that blinks occur at the completion of processing of a stimulus. Accordingly, measures of blink latency (relative to the presentation of some information) have been taken as an indication of processing demands. This is a possible explanation for instances in which blink rate increases (rather than decreases) with an increase in visual complexity (Wilson and Eggemeier, 1991).

To summarize, while measures of endogenous eye blink activity can be sensitive to variations in visual demand, it is not clear that these measures are sensitive to changes in auditory or cognitive demands (Wilson and Eggemeier, 1991).

In conclusion, physiological measures of workload have a number of characteristics that make them potentially useful in the context of evaluating operator performance in complex systems:

- Some physiological measures are relatively unobtrusive in that their implementation usually does not require activities that might change the operator's task.

- They can be recorded continuously, i.e., measured at the same time that operators are performing their tasks. Thus they do not rely on retrospective evaluations.

- The various physiological indicators have been shown to reflect different aspects of task demand.

Physiological measures can have disadvantages that can limit their usefulness in an operational testing situation:

- The instrumentation required to record physiological indicators may require the use of sensors or other devices attached to operators which may be uncomfortable and may not be acceptable to operators. Such equipment may also contribute to operators having the perception that the test environment is artificial.

- The small electrical signals on which some physiological measures are based can be obscured by electrical noise in the testing environment. Similarly, other electrophysiological signals (e.g., those associated with muscle activity) may degrade the signal of interest in situations where the operator is permitted to move about and speak to other operators.

- The instrumentation required to record and analyze physiological indicators and the expertise required to interpret the measures also make physiological measures of workload less favorable in some circumstances.

Overall Conclusions

Predictive techniques do not require operators to participate in simulated events. Thus, they are typically used in the early stages of design development and, therefore, may have limited application in integrated system validation. Empirical techniques are typically used during test and evaluation at later design stages and typically involve operators performing tasks on simulators. Because integrated system validation occurs late in the design process, empirical workload assessment techniques are more appropriate for workload assessment.

Of the techniques discussed, subjective and spare mental capacity may be the most suitable approaches. Both have been extensively used in system design and evaluation. Both can be implemented in ways to facilitate the determination of which aspects of workload are high.

Wierwille and Eggemeier (1993) identified a number of additional considerations for workload assessment. In evaluations of the performance of integrated systems, it is also desirable for the workload measure(s) to reflect short-term effects on operator workload. In operational relevant situations, task demands will vary from moment to moment. Short-term increases in demands may cause momentary overload which will not be reflected in measures taken over long intervals. Further, it is important to consider workload history, i.e., the levels of workload the operator has experienced prior to the time of interest (Smolensky and Hitchcock, 1993; Huey and Wickens, 1993). Decreases, as well as increases, in workload from accustomed levels may disrupt operators' performance. Accordingly, the capability to provide continuous measures of workload is an important characteristic of a potential workload measurement technique.

5.6.2.4 Anthropometric and Physiological Factors

Anthropometric and physiological factors include such concerns as visibility and audibility of indications, accessibility of control devices to operator reach and manipulation, and the design and arrangement of equipment. Many of these issues are the subject of evaluations conducted earlier in the design process. They may be included in validation activities as a check against unforeseen problems. Attention should be focussed on those anthropometric and physiological factors that can only be addressed during testing of the integrated system, e.g., the ability of the operators to effectively use the various controls, displays, workstations, or consoles in an integrated manner.

5.6.3 Performance Criteria

A performance measure only describes performance; it does not evaluate performance (ANSI, 1993). The goal of measurement is to allow a conclusion to be drawn regarding the system that is being validated, specifically with regard to its safety and support for effective operator interaction. In order to judge the acceptability of system performance, it is necessary to establish criteria for the performance measures used in the evaluations. Performance criteria are the standards against which the observed integrated system performance is compared to judge its acceptability.

There are several basic approaches to establishing criteria, which vary based upon the type of comparisons that are performed: requirement-referenced, benchmark referenced, normative referenced, and expert-judgement referenced.

Requirement Referenced

This is a comparison of the performance of the integrated system with respect to an accepted, quantified, performance requirement. For many variables a requirement-referenced approach can be used; i.e., requirements for plant, system, and operator performance can be defined through engineering analysis as part of the design process. Plant parameters governed by technical specifications and time requirements for critical operator actions are examples of performance measures for which a requirement-referenced criteria can be determined. For performance measures where such specific requirement referenced criteria cannot be used alternative criteria development methods must be used.

Benchmark Referenced

This is a comparison of the performance of the integrated system with that of a benchmark system which is predefined as acceptable under the same conditions or equivalent conditions. Such an approach is typically employed when no accepted independent performance requirements can be established. Performance is evaluated through comparisons to an accepted benchmark rather that through an absolute measurement. For example, the evaluation may test whether the plant under review can be operated to stay within a level of operator workload not exceeding that associated with Plant X. Plant X is identified as acceptable for reasons such as its acceptable operating history and operators report their workload levels to be acceptable. In this case the performance measure must be obtained for Plant X and the new system, under similar operational conditions, and then compared. In the establishment of benchmark-referenced criteria, similar test conditions should be established for the benchmark system and system under evaluation.

Normative Referenced

Normative-referenced comparison is similar to a benchmark reference comparison, however, the performance criterion is not based upon a single comparison system, it is based upon norms established for the performance measure through its use in many system evaluations. The new system performs as compared to the norms established under the same conditions or equivalent conditions. This approach can be used when no accepted independent performance requirements can be established, but repeated use of the same performance measure enables the development of performance norms for acceptable and unacceptable systems.

This has been done in other industries, e.g., the use of Cooper-Harper scale (Wierwille and Casali, 1983) and more recently the NASA-TLX (Hart and Staveland, 1988) are examples of this approach. The aerospace industry has established the meaning of these workload scales through their repeated application in numerous design evaluations. Designers could establish this type of criteria for NPP design. The advantage of this approach over benchmark criteria is that the measure can be used in the evaluation of different designs. The establishment of industry norms provides the individual designer with established criteria without the burden of identifying performance dimensions and then measuring reference systems. Due to the scope of such efforts, the development of normative-referenced criteria may be appropriate for an industry effort, rather than an individual designer.

Expert-Judgement Referenced

This is a comparison of the performance of the integrated system with criteria established through the judgement of SMEs.

Integrated system validation will require a combination of these approaches, since the types of performance to be measured are qualitatively different.

5.7 Test Design

Test design refers to the process of developing plans and conducting validation tests once the integrated system has been defined and measures have been selected. The goal of test design is to permit the observation of integrated system performance in a manner that avoids or minimizes bias, confounds, and noise (error variance). Shortcomings in test design can (1) alter the relationship between the integrated system and observations of performance, and/or (2) create enough noise to performance data to making results difficult to interpret. Such effects can compromise test design validity and thereby limit the generalizability of validation results to actual plant performance.

This section describes characteristics of the design of validation tests that are important to supporting test design validity. The following topics are addressed as subsections:

- Coupling Crews and Scenarios
- Test Procedures
- Test Conductor Training
- Participant Training
- Pilot Study.

5.7.1 Coupling Crews and Scenarios

The coupling of crews and scenarios refers to the process of determining how the test participants experience the test scenarios. It involves two steps. First is scenario assignment, the determination of which crews experience which scenarios. Second is scenario sequencing, the determination of the order in which each crew receives their scenarios. Each is discussed below.

Scenario Assignment

In research, the assignment of test participants to levels of the independent variables is referred to as the experimental design (Kirk, 1982). For example, assume there is an independent variable called "scenario" which has two levels - easy and difficult. A decision has to be made as to whether an individual participant will experience one or both levels of that variable. An independent variable is referred to as a within-subjects variable if every test participant is exposed to each level of the independent variable; i.e., both easy and difficult scenarios. A between-subjects variable is one in which every test participant is exposed to only one level of the independent variable; i.e., either the easy or difficult scenario but not both.

A given experiment can have a combination of independent variables. If all the variables are between-subjects, then each test participant is randomly assigned to only one of the test conditions (one

combination of the independent variables). This is called a factorial design. When all the variables are within-subjects, an individual test participant is assigned to all of the test conditions (all combinations of the independent variables). This is called a block design, where the participants represent blocks of data. This may also be called a repeated measures design (performance measures are repeated across test conditions using the same participants). When there is a combination of within and between subjects variables in one experiment the design is called a split-plot factorial.

Validation tests differ from the typical experiment in this regard in two ways. First, one is not generally interested in the effects of individual independent variables, rather, they are combined from scenarios. The validation team is interested to determine whether the integrated systems performance is acceptable under any and all scenarios. It is only when performance is unacceptable that the validation team may try to ascertain what variable may be responsible (see Section 5.8). Thus, in general, instead of assigning participants to levels of independent variables, it is more appropriate to think in terms of assigning crews to scenarios.

Second, there will in most cases be more scenarios than participant crews. This fact, in combination with the expense and effort of training crews, renders an opportunity to utilize randomized factorial designs impossible. Thus, crews will participate in more than one scenario. Where each crew can participate in each scenario, the design represents a repeated-measures design. However, there may be practical reasons why each crew will not be able to participate in all scenarios (due to factors such as crew availability or concerns over performance transfer from one scenario to another). In such cases a given crew will participate in some but not all scenarios. In research, this is called an incomplete block design.

Validation will typically involve either block or incomplete block design. When a complete block design is used, the next consideration is the sequence in which scenarios are presented (discussed in the next subsection below). When an incomplete block design is used, consideration should be given to balancing the set of scenarios so that each crew receives a representative range. This can be accomplished be using the operational event sampling dimensions, described in Section 5.5.1, to avoid confounding the performance of individual crews with the types of scenarios. For example, it would complicate the evaluation if Crew 1 received all the easy scenarios and Crew 2 received all the difficult scenarios. Suppose Crew 1 was a below average crew and Crew 2 was an above average crew. The data may indicate successful performance under all scenarios and one might be tempted to conclude that the design was validated. However, it is plausible that if the assignment of crews were reversed, such that Crew 1 received the difficult scenarios and Crew 2 the easy scenarios, the design would have been called into question because Crew 1 couldn't successfully operate the plant under difficult conditions. In this hypothetical example, the confounding of crews and scenarios would have led to a spurious validation of the design. Crew variability is a genuine phenomenon and its effect across the types of scenarios must be represented in order to appropriately test the design.

While an incomplete block always leads to partial confounding of the participants to scenarios, the negative effects can be greatly minimized by attempting to balance the important characteristics of scenarios across crews. It should be further noted that random assignment of scenarios to crews is *not* recommended. The value of using random assignment to control bias is only effective when the number of crews is quite large. Instead, the validation team should attempt to provide each crew with a similar and representative range of scenarios.

Scenario Sequencing

Another type of confounding that can occur is associated with sequence effects; i.e., effects caused by the order in which test scenarios are presented to the participants. Even when crews are trained to a performance criterion (see Section 5.7.4) prior to validation testing, they will become more experienced as the test proceeds and their performance may change. Their behavior may also systematically change for other reasons. One should attempt to prevent such changes from being confounded with the effects of scenarios. Thus, the order of presentation of scenario types to crews should be carefully balanced to ensure that the same types of scenario are not always being presented in the same linear position, e.g., the easy scenarios are not always presented first. There may also be subtle effects on performance of one scenario on another, e.g., something that happens in Scenario A may provide a clue for Scenario B (such as "I didn't think to look at the parameter X in the last scenario, so I will be sure to check it this time). For these reasons, it is desirable to not have Scenario A always follow scenario B.

The test design should establish an order of presentation of test scenarios for each crew that avoids these potential problems. There are formal approaches to this problem which may be applicable to a given validation program. For example, use of a Latin square arrangement of scenarios can control for sequence effects. Figure 5.1 illustrates a latin square arrangement of three scenarios (labeled A, B, and C) for each of three crews. Thus, for example, Crew 1 receives the scenarios in the order A, B, and C. However, such an approach may not always be practical, e.g., having at least as many crews as scenarios. In such cases the logic should be applied to arrange sequences to minimize the potential for sequence effects to confound the data.

SCENARIO ORDER	**CREW**		
	1	2	3
First	A	B	C
Second	B	C	A
Third	C	A	B

Figure 5.1 Latin square arrangement of three scenarios and three crews
(Scenarios are designated by the letters A, B, and C)

5.7.2 Test Procedures

Detailed, clear, and objective procedures should be available to govern the conduct of the tests. These procedures should include:

- Information pertaining to the experimental design, i.e., an identification of which crews receive which scenarios and the order that the scenarios should be presented.

- Detailed and standardized instructions for briefing the participants. The type of instructions given to participants can affect their performance on a task. This source of bias can be minimized by developing standard instructions.

- Specific criteria for the conduct of specific scenarios, such as when to start and stop scenarios, when events such as faults are introduced, and the other information discussed in Section 5.5.2, Scenario Definition.

- Scripted responses for test personnel who will be acting as plant personnel during test scenarios. To the greatest extent possible, responses to communications from operator participants to test personnel (serving as surrogate outside the control room personnel) should be prepared. There are limits to the ability to preplan communications since operators may ask questions or make requests that were not anticipated. However, efforts should be made to detail what information personnel outside the control room can provide, and script the responses to likely questions.

- Guidance on when and how to interact with participants when simulator or testing difficulties occur. Even when a high-fidelity simulator is used, the participants may encounter artifacts of the test environment that detract from the performance for tasks that are the focus of the evaluation. Guidance should be available to test conductors to help resolve such conditions.

- Instructions regarding when and how to collect and store data. These instructions should identify which data are to be recorded by:

  - simulation computers,
  - special purpose data collection devices (such as automated situation awareness data collection, workload measurement, or physiological measures),
  - video recorders (locations and views),
  - test personnel in real time (such as observation checklists), and
  - subjective rating scales and questionnaires.

- Instructions for maintaining and updating test conductor logs. These instructions should detail the types of information that should be logged (e.g., when tests were performed, deviations from test procedures, and any unusual events that may be of importance to understanding how a test was run or interpreting test results) and when it should be recorded.

- Procedures for documentation, i.e., identifying and maintaining test record files including crew and scenario details, data collected, and test conductor logs.

Where possible the use of a double-blind procedure should be used to minimize the opportunity of tester expectancy bias or participant response bias (see Section 4.2.4, Test Design Validity, for a discussion of these potential sources of bias) in response to demand characteristics. A double-blind procedure is one in which neither the operator participants nor the test personnel who directly interact with them know any details of the scenario to be conducted.

5.7.3 Test Conductor Training

Test conductor personnel are those members of the validation team who will actually conduct the validation tests. These personnel should be trained on the use and importance of test procedures. This training should address experimenter bias and the types of errors that may be introduced into test data through the failure of test conductors to accurately follow test procedures or interact properly with participants. The importance of accurately documenting problems that arise in the course of testing, even if due to test conductor oversight or error, should be emphasized. Failure to note such problems could result in misleading and even incorrect conclusions regarding the acceptability of integrated system performance.

5.7.4 Participant Training

Participant training is an essential part of validation and should be of high fidelity; i.e., highly similar to that which personnel will receive in an actual plant. The participants should be trained to ensure that their knowledge of the operator's role, concept of operations, the plant design, and use of the HSI is representative of anticipated users of the plant. This will help assure that the participants are representative of actual users. It may be possible to limit training to the scope of the validation tests, however, participants should not be trained specifically to perform the validation scenarios. Failure to appropriately train participants is a potential threat to the validity of the study. It can create bias and increased noise. If training is different than from that which actual plant personnel will receive, then the generalizability of the validation test results to actual plant performance may be threatened.

Training is important for two reasons. First, inadequate participant training may result in poor performance and, consequently, negatively biased evaluations of the design. Second, incomplete or inadequate participant training may result in the test results being affected by learning effects on the part of the participants. Learning effects typically reflect a high rate of improvement in the early trials followed by a decreasing rate of improvement in the later trials. The point at which performance no longer improves with continued practice is called asymptotic performance. Unless participants are trained essentially to asymptotic performance before the test trials begin, test data will reflect the learning process. Whether these effects represent a confound or negatively bias performance will depend upon the experimental design considerations discussed above. Therefore, participants should be trained and tested prior to conducting actual test trials. Participants should be trained to a performance criteria similar to that which will be applied to actual plant personnel.

5.7.5 Pilot Testing

A pilot study should be conducted prior to conducting the integrated validation tests to provide an opportunity to assess the adequacy of the test design, performance measures, and data collection methods (ANSI, 1993; Conrad and Maul, 1981; Meister, 1986; Muckler and Stevens, 1992). Aspects to the test that are found to be infeasible can be changed prior to conducting the full validation test. Pilot studies also provide an opportunity to estimate important performance measurement parameters, such as variability. These estimates can be used to assess the degree to which decisions can be drawn to test results (Muckler and Stevens, 1992).

Personnel who will participate in the validation tests should not participate in the pilot study. If the pilot study is conducted using the validation test participants then:

- The scenarios used for the pilot study should be different from those used in the validation tests, and

- Care should be given to ensure that the participants do not become so familiar with the data collection process that it may result in response bias (Conrad and Maul, 1981).

5.8 Data Analysis and Interpretation

As was discussed in Section 4.2.4, Statistical Conclusion Validity, performance measures should be examined with respect to:

- The relationship between the performance data and the established performance criteria, and

- The inference from observed performance to estimated "population" performance.

Similar considerations are addressed in numerous standards as well. The Draft IEC standard (IEC, in preparation) indicates that ample margins should be observed in performance measures, such as task times, to account for human variability and recommends the use of statistical analysis. Similarly, the ANSI standard (1993) indicates that "inferential and descriptive statistics express HPM (human performance measurement) data in terms of the population in a manner that encourages confidence in their accuracy and generalizability….an inferential statistical test should demonstrate that the results or conclusions have less than a 5% probability of having occurred by chance" (pp. 31-32).

However, several factors combine to make a statistical analysis such as that obtained from research data difficult to perform for integrated system validation. First, because of the need to test the integrated system under a wide range of operational conditions, there may not be sufficient data under one set of constant conditions to provide reliable estimates of population performance parameters. As indicated in Table 3.1, this is one of the significant differences between validation and research. Second, one may not be able to think in terms of deviations from an optimal or mean performance because of operator strategy differences, i.e., because there may be no single strategy that is required the mean may not be a meaningful parameter. Therefore, validation data should be analyzed through a combination of quantitative and qualitative methods. The analysis should consider the potential for Type 1 errors, i.e., concluding that the design is acceptable when actual performance is unacceptable (incorrectly validating the design); and Type 2 errors, i.e., concluding that the design is unacceptable when actual performance is acceptable (incorrectly rejecting the design).

For all performance measures, descriptive statistics such as measures of central tendency and variability should be provided and compared to performance criteria. The specific measures used should be appropriate to the level of measurement of the performance measure (e.g., it would be inappropriate to report a mean for ordinal scale data). Where possible, inferential statistics (Keppel, 1982; Kirk, 1982)should be calculated to determine whether observed performance is reliably within acceptable performance envelopes. For nonparametric data, non-parametric tests of significance should be employed (Siegel and Castellan, 1988). For all analyses, statistical parameters and tests should be appropriate to the measurement scale of the performance measure, e.g., nominal, ordinal, interval, or ratio scale of measurement.

The analyses of test data should be independently verified for correctness. There is a tendency to more carefully recheck results that are not favorable which is a form of experimenter bias. However, any result can be subject to error, thus verification is a necessary check. All raw data and the formulas used for their analysis should be documented for independent review.

Where the statistical assumptions cannot justify the use of statistical tests or where the sample size for a desired comparison is too small, qualitative comparisons of the observed variability in performance and the performance criteria should be made to determine whether sufficient margin exists to permit prediction of successful performance in the actual system. The basis for the determination should be clearly documented.

The degree of convergence of the multiple measures of performance should be evaluated. When all the measures of performance are considered, there should be consistency of statistical conclusions. Where performance is acceptable on some measures and unacceptable on others, further analysis is warranted. Once an instance of unacceptable performance has been identified, consideration should be given to its cause; possible root causes include: a design deficiency in the integrated system, an artifact of the testing process, or inadequate sample size. If unacceptable performance reflects an artifact of the testing process, and the deficiency corrupts the inference process, then the test methodology should be revised and the tests should be repeated. If unacceptable performance reflects uncertainties in estimates of actual performance due to high variance (low statistical power), then additional data should be collected to provide more reliable performance estimates.

If the unacceptable performance is due to a design deficiency, consideration should be given to its root cause. It is important to maintain the perspective that the unit of analysis is the integrated system. Thus, deficiencies can be the result of problems with any of the constituent parts or their interactions; e.g.:

- Function allocation (inappropriate use of automation),
- Task definition (failure to properly identify the information, decision, control, and feedback requirements),
- Staffing/job design (poor allocation of tasks to personnel, deficiencies in crew coordination and communication),
- Training (training program failures to prepare personnel for operations),
- HSI (failure to properly consider human performance tradeoffs in HSI component selection; inappropriate allocation of HSI functional requirements to HSI components such as group-view displays and workstation displays; deficiencies in the design of alarms, displays, controls, job aids, and procedures; poorly designed user interface management; failure to consider human performance effects of extreme environments).

To help analyze human performance problems, the dimensions that were combined to develop the problematic scenarios should be reviewed. This is essentially a process of tracing back to the "independent variables" of the operational event selection process (see Section 4.5). If the sampling process had successfully identified the plant and operational characteristics that contribute to the

variability of system performance, examining the specific dimensions that make up problematic scenarios should contribute to the identification of the root cause of performance problems.

It is, of course, possible that performance problems were due to a unique interaction between important dimensions. Such interactions may be much more difficult to detect unless the same interactions lead to performance issues in multiple scenarios.

Each deficiency should be considered with regard to its impact on plant safety. As per NUREG-0700, Rev. 1 (O'Hara, et al., 1995), the potential effects of these deficiencies should be determined, in part, by the safety significance of the plant system(s) impacted, the safety significance of the personnel function (e.g., consequences of failure), the effect on SAR accident analyses, and their relationship to risk significant sequences in the plant PRA. Deficiencies identified as having significant safety consequences are those in which the consequences of personnel error could reduce the margin of plant safety below an acceptable level, as indicated by such conditions as violations of technical specification safety limits, operating limits, or limiting conditions for operations.

Deficiencies should be prioritized as follows. First priority deficiencies should be those with direct safety consequences and those with indirect or potential safety consequences. Deficiencies with direct safety consequences include violations of personnel information requirements for personnel tasks that are related to plant safety. Deficiencies with indirect safety consequences include those that would seriously affect the ability of personnel to perform the task. The severity of the deficiency should be assessed in terms of the degree to which it contributes to human performance problems such as workload and information overload.

Second priority deficiencies should be those that do not have significant safety consequences, but do have potential consequences to plant performance/operability, non-safety-related personnel performance and efficiency, or other factors affecting overall plant operability. These include personnel tasks associated with plant productivity, availability, and protection of investment. The remaining deficiencies should be those that have little or no consequence to plant safety or operation.

Each deficiency should be fully documented including: priority, associated plant system, associated integrated system component (as identified above), and associated personnel function. The documentation should clearly indicate whether the deficiency was dismissed or identified as in need of design modification, and the basis for this determination in terms of consequence to plant safety or operation should be clearly described.

Design solutions should be identified to address deficiencies. Where deficiencies are determined to be of minimal safety significance and where the causes are understood, design changes may be subject to limited, focused testing. If deficiencies have greater significance or the causes are not well understood, then integrated system validation tests should be repeated. Special attention should be given to the inter-relationship of many individual design modifications. When it is not possible to fully correct the problems identified by an deficiency, justification should be provided.

5.9 Validation Conclusions

Following the analysis of data and resolution of any issues as discussed in the previous section, conclusions should be drawn with regard to integrated system validation. In Section 4.3, the

characteristics of a validated system were presented. These characteristics should be considered with respect to the entire integrated system validation program.

The integrated system may be considered validated if the following is demonstrated (see Section 4.3 for a further description of the types of validity and their major considerations and threats):

1. A comprehensive testing program was conducted by an independent, multidisciplinary team.

2. System representation validity is logically supported such that the integrated systems is concluded to be representative of the actual system in all aspects that are important to integrated system performance. Constant aspects of the system, model and HSI, are high-fidelity and variable aspects of the system were adequately sampled and represented in high-fidelity. The major threats to system representation validity are ruled out, including:

    - Inadequate process/plant model fidelity
    - Inadequate HSI fidelity
    - Inadequate participant fidelity
    - Participant sampling bias
    - Historical population changes
    - Operational conditions sampling bias
    - Inadequate scenario fidelity

3. Performance representation validity is logically supported such that the measures of integrated system performance and their associated criteria reflect good measurement practices and are concluded to be representative of important aspects of performance. The major threats to performance representation validity are ruled out, including:

    - Test-level underspecification
    - Measurement underspecification
    - Changing measures
    - Poor measurement characteristics
    - Underspecified performance criteria
    - Measurement-scenario interaction.

4. Test design validity is logically supported such that there are no plausible biasing or confounding effects to make the predictions of system performance ambiguous. The major threats to test procedure validity are ruled out, including:

   - Test procedure underspecification bias
   - Tester expectancy bias
   - Participant response bias
   - Test environment bias
   - Changes in participants over time
   - Participant assignment bias
   - Sequence effects

5. Statistical conclusion validity is logically supported and based upon a convergence of the multiple measures such that it can be concluded that the performance of the actual system will be acceptable. The major threats to statistical conclusion validity are ruled out, including:

   - Accepting narrow performance margins
   - High noise
   - Low sample size

When these conditions are met, the results of the validation process is considered to be representative of the actual system performance and generalization is supported. In essence, the validation test program has failed to invalidate the design.

6.0 REFERENCES

Adams, M., Tenney, Y., and Pew, R., "Situation Awareness and the Cognitive Management of Complex Systems," *Human Factors*, 37:85-104, 1995.

Alessi, S. M., "Fidelity in the Design of Instructional Simulations," *Journal of Computer-Based Instruction,* 15, 2:40-47, 1988.

American National Standards Institute, ANSI/ANS-58.8-1994, "American National Standard Time Response Design Criteria for Safety-Related Operator Actions," La Grange Park, IL., 1994.

American National Standards Institute, ANSI/AIAA G-035-1992, "Guide to Human Performance Measurements," Washington, DC., 1993.

American Nuclear Society, ANSI/ANS-3.5-1985, "Nuclear Power Plant Simulators for Use in Operator Training," La Grange Park, IL., 1985.

Bainbridge, L., "What Should a Good Model of the NPP Operator Contain," *Proceedings of the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems*, American Nuclear Society: IL, 1986.

Bainbridge, L., "Analysis of Verbal Protocol from a Process Control Task," *The Human Operator in Process Control*, Taylor and Francis: London, England, 1974.

Baker, S. and Marsall, E., "Evaluating the Man-Machine Interface - The Search for Data", (J. Patrick and K. D. Duncan, eds.), *Training, Human Decision Making and Control*, 79-92, New Yok: Elsevier Science, 1988.

Barriere, M. and Luckas, W., *A technique for human error analysis (ATHEANA) - An initial demonstration* (Draft NUREG/CR), U.S. Nuclear Regulatory Commission, Washington, D.C., in preparation.

Beare, A.N. and Dorris, R.E., "The Effects of Supervisor Experience and the Presence of a Shift Technical Advisor on the Performance of Two-Man Crews in a Nuclear Power Plant Simulator," *Proceedings of the Human Factors Society - 28nd Annual Meeting*, Santa Monica, California: Human Factors Society, 1984.

Bittner, A.C., "Robust Testing and Evaluation of Systems: Framework, Approaches, and Illustrative Tools," *Human Factors*, 34:477-484, 1992.

Campbell, D. and Fisk, D., "Convergent and Discriminant Validation by the Multitrait-multimethod Matrix," *Psychological Bulletin*, 56:81-105, 1959.

Chapanis, A. and VanCott, H., "Human Engineering Tests and Evaluations," (H. VanCott and R. Kinkade, eds.), *Human Engineering Guide to Equipment Design*, US Government Printing Office, Washington, DC., 1972.

Chubb, G.P., Laughery, R.K., and Pritsker, A.A.B., "Simulating Manned Systems," (G. Salvendy, ed.), *Handbook of Human Factors*, New York: Wiley, 1987.
Cohen, F., *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York, 1969.

Conrad, E. and Maul, T., *Introduction to Experimental Psychology*, John Wiley and Sons, New York, 1981.

Cook, T. and Campbell, D., *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Co., Boston, MA., 1979.

Echeuerria, D. et al., *The Impact of Environmental Conditions on Human Performance*, (NUREG/CR-5680) U.S. Nuclear Regulatory Commission, Washington, D.C., 1994.

Eggemeier, F.T. and Wilson, G.F., "Performance-Based and Subjective Assessment of Workload in Multi-Task Environments," (Damos, D.L., ed.), *Multiple-Task Performance*, Taylor and Francis London, 1991.

Electric Power Research Institute, *Advanced Light Water Reactor Utility Requirements Document,* Volume II, *Evolutionary Plant,* Revision 4, Palo Alto, California, 1992.

Endsley, M., "Design and Evaluation for Situation Awareness Enhancement," *Proceedings of the Human Factors - 32nd Annual Meeting*, Human Factors Society, California, 1988.

Endsley, M.R., "Predictive Utility of an Objective Measure of Situation Awareness," *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, 1990.

Endsley, M.R., "Situation Awareness in Dynamic Human Decision Making: Measurement," *Proceedings of the 1st International Conference on Situation Awareness in Complex Systems*, 1993a.

Endsley, M.R., "Situation Awareness and Workload: Flip Sides of the Same Coin," *Proceedings of the 7th International Symposium on Aviation Psychology*, 1993b.

Endsley, M., "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, 37:32-64, 1995a.

Endsley, M., "Measurement of Situation Awareness in Dynamic Systems," *Human Factors*, 37:65-84, 1995b.

Fraker, M., "A Theory of Situation Awareness: Implications for Measuring Situation Awareness," *Proceedings of the Human Factors Society - 32nd Annual Meeting*, Human Factors Society, Santa Monica, California, 1988.

Gartner, W. and Murphy, M., "Pilot Workload and Fatigue: A Critical Survey of Concepts and Assessment Techniques," (NASA TN D-8365), National Aeronautical and Space Administration, Washington, D.C., 1976.

Gopher, D. and Donchin, E., "Workload: An Examination of the Concept," (K.R. Boff, L. Kaufman, and J. Thomas, eds.), *Human of Perception and Human Performance, Vol. 2. Cognitive Processes and Performance*, Wiley, New York, 1986.

Gould, J., "How to Design Usable Systems," (M. Helander, ed.), *Handbook of Human-Computer Interaction*, Elsevier Science Publishers, Amsterdam, 1988.

Hancock, P. and Meshkati, N., *Human Mental Workload*, North-Halland, New York, 1988.

Hart, S. and Staveland, L., "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," (P. Hancock and N. Meshkati, eds.), *Human Mental Workload*, North-Holland, New York, 1990.

Hart, S. and Wickens, C., "Workload Assessment and Prediction," (H. Booher, ed.), *Manprint: An Approach to Systems Integration*, Van Nostrand Reinhold, New York, 1990.

Hicks, T. and Wierwille, W., "Comparison of Five Mental Workload Assessment Procedures in a Moving-Base Driving Simulator," *Human Factors*, 21:129-143, 1979.

Hogg, D., Folleso, K., Torralba, B., and Volden, F., *Measurement of the Operator's Situation Awareness for Use Within Process Control Research: Four Methodoligical Studies, (HWR-377)*, OECD Halden Reactor Project, Halden, Norway, 1994.

Hollnagel, E., "The Reliability of Interactive Systems: Sumulation Based Assessment," (J. Wise, D. Hopkin, D., and P. Stager, P., eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1993.

Huey, B. and Wickens, C., "Workload Transition: Implications for Individual and Team Performance," National Academy Press, Washington, D.C., 1993.

Institute of Electrical and Electronics Engineers, IEEE Standard 1023-1988, "IEEE Guide to the Application of Human Factors Engineering to Systems, Equipment, and Facilities of Nuclear Power Generating Stations," New York, 1988.

International Atomic Energy Agency (IAEA), International Nuclear Safety Advisory Group, Safety Series No. 75-INSAG-3, "Basic Safety Principles for Nuclear Power Plants," Vienna, Austria, 1988.

International Electrotechnical Commission (IEC), "Verification and Validation of Control Room Design of Nuclear Power Plants," IEC-Draft Standard (5th), Bureau Central de la Commission Electrotechnique Internationale, Geneva, Switzerland, in preparation.

Jex, H., "Measuring Mental Workload: Problems, Progress, and Promises," (P. Hancock and N. Meshkati, eds.), *Human Mental Workload*, North-Halland, New York, 1988.

Kantowitz, B. H., "Selecting Measures for Human Factors Research," *Human Factors*, 34:387-398, 1992.

Kantowitz, B.H., "Can Cognitive Theory Guide Human Factors Measurement?", *Proceedings of the Human Factors Society 34th Annual Meeting*, Santa Monica, 1990.

Keppel, G., *Design and Analysis*, second edition, Prentice Hall, Englewood Cliffs, NJ., 1982.

Kirk, R., *Experimental design*, second edition, Brooks/Cole. Pub. Co., Belmont, CA, 1982.

Knowles, W.B., "Operator Loading Tasks," *Human Factors*, 5:155-161, 1963. (reprinted in M. Venturino, ed.), *Selected Readings in Human Factors*, Santa Monica, 1990.

Kramer, A.F., "Physiological Metrics of Mental Workload: A Review of Recent Progress," (Damos, D.L., ed.), *Multiple-Task Performance*, Taylor and Francis, London, 1991.

Kramer, H. and Thiemann, S., *How Many Subjects?: Stastical Power Analysis in Research*, Sage Publications, London, 1987.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., Wherry, R.J., "Operator Workload: Comprehensive Review and Evalution of Operator Workload Methodologies (Tech Rep 851)," U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA., 1989.

Meister, D., *Conceptual Aspects of Human Factors*, The Johns Hopkins University Press, Baltimore, 1989.

Meister, D., *Human Factors in Testing and Evaluation*, Elsevier, Amsterdam, 1986.

Meister, D., *Behavioral Analysis and Measurement Methods*, Wiley, New York, 1985.

Meshkati, M. and Lowenthal, A., "The Effects of Individual Differences in Information Processing Behavior in Experiencing Mental Workload and Perceived Task Difficulty: A Preliminary Investigation," (P. Hancock and N. Meshkati, eds.), *Human Mental Workload*, North-Halland, New York, 1988.

Miller, D., Wolfman, J., Mullins, R., and Crehan, C., "Beyond the Bounds of the Human Factors Tool Kit: Computer-Human Interface Design in a Complex System," *Proceedings of the Human Factors Society 38th Annual Meeting*, Human Factors Society, Santa Monica, CA., 1994.

Moray, N., "Monitoring Behavior and Supervisory Control," *Handbook of Human Perception and Performance*, John Wiley and Sons: New York, NY, 1986.

Moray, N., *Mental Workload: Its Theory and Measurement*, Plenum Press, NY, 1979.

Moray, N., Lootsteen, P., and Pajak, J., "Acquisition of Process Control Skills," *IEEE Transactions on Systems, Man, and Cybernetics*, 16:497-504, 1986.

Muckler, F.A. and Stevens, S.A., "Selecting Performance Measures: "Objective" Versus Subjective Measurement, *Human Factors*, 43:441-455, 1992.

Norman, D., *The Psychology of Everyday Things*, Basic Books, New York, 1988.

Norman, D., "Categorization of Action Slips," *Psychological Review*, 88:1-15, 1981.

Nunnally, J., *Psychometric Theory*, McGraw Hill, New York, 1967.

O'Donnell, R.D. and Eggemeier, F.T., "Workload Assessment Methodology," (Boff, K.R., Kaufman, L., and Thomas, J.P.), *Handbook of Perception and Human Performance*, Wiley, New York, 1986..

O'Hara, J., *Advanced Human System Interface Design Review Guideline: General Evaluation Model, Technical Development, and Guideline Description*, (NUREG/CR-5908, Volume 1), U.S. Nuclear Regulatory Commission, Washington, D.C., 1994.

O'Hara, J., Brown, W., Stubler, W., Wachtel, J. and Persensky, J., "Human-System Interface Design Review Guideline," (Draft NUREG-0700, Rev. 1), U.S. Nuclear Regulatory Commission, Washington, D.C., 1995.

O'Hara, J., Higgins, J., Stubler, W., Goodman, C., Eckenrode, R., Bongarra, J., and Galletti, G., *Human Factors Engineering Program Review Model*, (NUREG-0711), U.S. Nuclear Regulatory Commission, Washington, D.C., 1994.

Patrick, J., "Information at the Human-Machine Interface," *New Technology and Human Error*, John Wiley and Sons: New York, NY, 1987.

Perrow, C., *Normal Accidents*, Basic Books, New York, 1984.

Popper, K., *The Logic of Scientific Discovery*, Basic Books, New York, 1959.

Rasmussen, J., *Coping Safely with Complex Systems*, Paper presented to the American Association for the Advancement of Science, Boston, 1988.

Rasmussen, J., *Information Processing and Human-Machine Interaction*, North Holland: New York, NY, 1986.

Rasmussen, J., "Skills, Rules, Knowledge: Signal, Signs, and Symbols and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems, Man, and Cybernetics*, 13:257-267, 1983.

Reason, J., *Human Error*, Cambridge University Press, New York, 1990.

Reason, J., "Modelling the Basic Error Tendencies of Human Operators," *Reliability Engineering and System Safety*, 22:137-153, 1988.

Reason, J., "Generic Error-Modelling Systems (GEMS): A Cognitive Framework for Locating Common Human Error Forms," *New Technology and Human Error*, J. Wiley and Sons: New York, NY, 1987.

Reid, G. and Nygren, T., "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload," (P. Hancock and N. Meshkati, eds.), *Human Mental Workload*, North-Holland, New York, 1988.

Reid, G.B., Shingledecker, C.A., and Eggemeier, F.T., "Application of Conjoint Measurement to Workload Scale Development," *Proceedings of the Human Factors Society - 25th Annual Meeting*, Santa Monica, 1981.

Reiersen, C. S., Baker, S., and Marsall, E., "An Experimental Evaluation of an Advanced Alarm System for Nuclear Power Plants - A Comparative Study," (J. Patrick and K. D. Duncan, eds.), *Training, Human Decision Making and Control*, 43-78, Elsevier Science, New York, 1988.

Rosness, R., "Limits to Analysis and Verification," (J. Wise, D. Hopkin, D., and P. Stager, P., eds.) *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1993.

Rouse, W. B., "A Mixed Fidelity Approach to Technical Training," *Journal of Educational Technology Systems,* 12(2):103-115, 1982.

Saaty, *The Analytic Hierarchy Process*, McGraw-Hill, New York, 1980.

Sagan, S., *The Limits of Safety*, Princeton University Press, New Jersey, 1993.

Sarter, N.B. and Woods, D.D., "Situation Awareness: A Critical but Ill-defined Phenomenon," *International Journal of Aviation Psychology*, 1:45-57, 1991.

Sheridan, T., "A General Model of Supervisory Control," *Monitoring Behavior and Supervisory Control*, Plenum Press: New York, NY, 1976.

Siegel, S. and Castellan, N. J., "Nonparametric Statistics for the Behavioral Sciences (2nd Ed.)," McGraw-Hill, 1988.

Smolensky, M.W. and Hitchcock, L., "When Task Demand is Variable: Verifying and Validating Mental Workload in Complex, "Real World" Systems," (J.A. Wise, V.D. Hopkin, and P. Stager, eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, Springer-Verlag, Bonn, 1993.

Stager, P., "Validation as a Means to Certification," *Proceedings of the Human Factors Society 38th Annual Meeting*, Santa Monica, CA.. 1994.

Stubler, W.F., Roth, E. M., and Mumaw, R., "Integrating Verification and Validation with the Design of Complex Man-machine Systems," (J. Wise, D. Hopkin, D., and P. Stager, P., eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1992.

Taylor, R.M., "Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design," *Situational Awareness in Aerospace Operations*, (AGARD Proceedings No. 478), 1989.

U.S. Department of Defense (DOD), "Human Engineering Requirements for Military Systems, Equipment and Facilities," MIL-H-46855B, Office of Management and Budget, Washington, D.C., 1979.

U.S. Nuclear Regulatory Commission, Regulatory Guide 1.149, "Nuclear Power Plant Simulation Facilities for Use in Operator License Examinations," Revision 2, Washington, D.C., 1987a.

U.S. Nuclear Regulatory Commission, *BWR and PWR Off-Normal Event Descriptions* (NUREG-1291), Washington, D.C., 1987b.

U.S. Nuclear Regulatory Commission, *Loss of Main and Auxiliary Feedwater Event at the Davis-Besse Plant on June 9, 1985*, (NUREG-1154), Washington, D.C., 1985.

Vidulich, M.A., "The Use of Judgement Matrices in Subjective Workload Assessment: The Subjective Workload Dominance (SWORD) Technique," *Proceedings of the Human Factors Society - 32nd Annual Meeting*, Santa Monica, 1988.

Vidulich, M.A., "The Cognitive Psychology of Subjective Mental Workload," (P. Hancock and N. Meshkati, eds.), *Human Mental Workload*, Elsevier, Amsterdam 1988.

Vidulich, M.A. and Hughes, E.R., "Testing a Subjective Metric of Situation Awareness," *Proceedings of the Human Factors Society 35th Annual Meeting*, 1991.

Vidulich, M.A., Ward, G.F., and Schueren, J., "Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment," *Human Factors*, 33:677-691, 1991.

Vincente, K., "Multilevel Interfaces for Power Plant Control Rooms I: An Integrative Review," *Nuclear Safety*, 33:381-397, 1992.

Vincente, K., "Supporting Knowledge-based Behavior Through Ecological Interface Design," EPRL-91-01, Engineering Psychology Research Laboratory at University of Illinois: Urbana, IL, 1991.

Vreuls, D. and Obermayer, "Human-system Performance Measurement in Training Simulators," *Human Factors,* 27:241-250, 1985.

Webb, E., Campbell, D., Schwartz, R., and Sechrest, L., *Unobtrusive Measures*, Rand McNally and Co., Chicago, 1973.

Wickens, C., *Workload and Situation Awareness: An Anology of History and Implications*, Insight, 94:1-3, 1992.

Wickens, C., "The Effects of Control Dynamics on Performance," (K. Boff, L. Kaufman, and J. Thomas, eds.), *Handbook of Perception and Human Performance: Volume II - Cognitive Processes and Performance*, John Wiley and Sons, New York, 1986.

Wickens, C., *Engineering Psychology and Human Performance,* Merrill Publishing Company, Columbus, Ohio, 1984.

Wickens, C., "The Structure of Attentional Resources," (R. Nickerson, ed.), *Attention and Performance VIII*, Erlbaum, NJ, 1980.

Wickens, C. and Kessel C., "The Detection of Dynamic System Failures," *Human Detection and Diagnosis of System Failures*, Plenum Press, New York, 1981.

Wieringa, P. and Stassen, H., "Assessment of Complexity," (J. Wise, D. Hopkin, D., and P. Stager, P., eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1993.

Wierwille, W. and Casali, J., "A Validated Rating Scale for Global Mental Workload Measurement Applications," *Proceedings of the 27th Annual Meeting of the Human Factors Society*, Santa Monica, CA, 1983.

Wierwille, W.W. and Eggemeier, F.T., "Recommendations for Mental Workload Measurement in a Test and Evaluation Environment," *Human Factors*, 35:263-281, 1993.

Wierwille, W. and Williges, R., *Survey and Analysis of Operator Workload Assessment Techniques*, (TR 2-78-101), Systemetrics, Corp., Blacksburg, VA., 1978.

Williges, R, and Wierwille, W., "Behavioral Measures of Aircrew Mental Workload," *Human Factors*, 21:549-574, 1979.

Wilson, G.F. and Eggemeier, F.T., "Psychophysiological Assessment of Workload in Multi-task Environments," (Damos, D.L., ed.), *Multiple-Task Performance*, Taylor and Francis, London, 1991.

Wise, J., Hopkin, D., and Stager, P., *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1993.

Wise, J., Hopkin, D., Stager, P., and Harwood, K., "Human Factors Ceritication of Systems," *Proceedings of the Human Factors Society 38th Annual Meeting*, Santa Monica, CA., 1994.

Woods, D. and Sarter, N., "Evaluating the Impact of New Technology on Human-machine Cooperation," (J. Wise, D. Hopkin, D., and P. Stager, P., eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, (NATO ASI Series F, Vol. 110), Springer-Verlag, Berlin, 1993.

Woods, D., and Sarter, N., "How in the World Did We Ever Get Into That Mode?, Mode, Error and Awareness in Supervisory Control," *Human Factors*, 37:5-19, 1995.

Woods, D., Johannesen, L., Cook, R., and Sarter, N., *Behind Human Error: Cognitive Systems, Computers, and Hindsight*, (CSERIAC SOAR 94-01), Crew Systems Ergonimics Information Analysis Center, Wright Patterson Air Force Base, Ohio, 1994.

Woods, D., O'Brien, J., and Hanes, L., "Human Factors Challenges in Process Control: The Case of Nuclear Power Plants," (G. Salvendy, ed.), *Handbook of Human Factors*, J. Wiley and Sons, New York, 1987.